UNIVERSITETI POLITEKNIK I TIRANES

METODA STATISTIKORE DHE ZBATIMET E TYRE

"Ndërtimi dhe analiza e një modeli probabilitaro - statistikor për studimin e efektit të ndotjes në gjendjen shëndetësore të banorëve në zona industriale"

M.Sc. Etleva Beliu (Llagami)

Udhëheqës Shkencor: Prof. Dr. Shpëtim Leka

Tirane 2015

REPUBLIKA E SHQIPERISE
UNIVERSITETI POLITEKNIK – TIRANE
FAKULTETI INXHINIERISE MATEMATIKE & INXHINIERISE FIZIKE
Departamenti i Inxhinierise matematike

DISERTACION

i paraqitur nga

M.Sc. Etleva Beliu (Llagami)

për marrjen e gradës shkencore

"DOKTOR"

në Inxhinieri Matematike

Tema: Metoda statistikore dhe zbatimet e tyre

"Ndërtimi dhe analiza e një modeli probabilitaro - statistikor për studimin e efektit të ndotjes në gjendjen shëndetësore të banorëve në zona industriale"

Udhëheqës Shkencor: Prof. Dr. Shpëtim Leka

Mbrohet më datë _____para jurisë

1._____ Kryetar

2._____ Anëtar

3._____ Anëtar

4._____ Anëtar

5._____ Anëtar

Abstrakti

Në këtë punim synohet që, me paraqitjen e metodave të ndryshme statistikore, analizën e detajuar të menyres së realizimit të tyre, t'i tregohet kujdes kushteve që ato kërkojnë për zbatim, dhe për më tepër, analizës së vlefshmërisë së tyre, apo si të thuash vlerësimin që modeli i ndërtuar në bazë të të dhënave të zgjedhjes, të mund të përdoret për tërë popullimin.

Kështu, gjatë studimit, synohet që krahas të përbashkëtave të metodave statistikore, të vërehen të vecantat e tyre, duke berë të mundur zgjedhjen e metodës më të përshtatshme në nje situata konkrete. Të gjitha këto metoda janë përdorur për ndërtimin dhe analizën e një modeli probabilitaro- statistikor për studimin e efektit të ndotjes në gjendjen shëndetësore të banorëve në zona industriale.

Është paraqitur e përdorur metoda e njohur ANOVA, por jo vetëm si një metodë e krahasimit të mesatareve, por janë interpretuar vlerësimet intervalore me të cilat ajo është shoqëruar, dhe përdorimi i saj i gjere në përfundimet e analizave të tërë metodave të tjera. Metoda e Regresit Linear te Shumëfishtë është parë në tërë llojshmeritë e tij, jo vetëm të regresit tëzakonshëm, por edhe të atij Hirarkik dhe Stepwise, si metoda që i japin përgjigje pyetjeve të ndryshme që lindin gjate modelimeve statistikore. Regresi Logjistik, si metoda më e mirë e analizes së nje modeli te nje ndryshoreje rasti kategorike, është përdorur, e madje është ndertuar edhe modeli probabilitaro-statistikor per gjendjen e banoreve të zonës në pneumologji. MANOVA,si metodë e krahasimit të mesatareve të komponentëve të një vektori është aplikuar, e përfundimet e saj janë krahasuar me rezultatet me të mira te metodës Analizës se Komponentëve Kryesorë, në interpretimin të gjendjes së gjendjes së përgjithshme të banorëve të zonës.

Gershetimi i terësisë së këtyre metodave ndihmoi edhe ne analizen komplekse të e problemeve banoreve në hematologji.

HYRJE

Analiza matematikore e modeleve statistikore, të kuptuarit në thellësi të të gjithë specifikave të tyre, njohja e hipotezave me të cilat këto metoda shoqërohen, statistikave që ato përdorin për vlerësim, vecantive të tyre, është hapi i pare në procesin e ndertimit te modeleve probabilitare-statistikore. Pa dyshim zhvillimi i teknologjisë kompjuterike dhe rritja e fuqisë dhe shpejtësisë së tyre përpunuese, ka rritur efektet e këtyre metodave në mënyrë dramatike

Këto metoda jane përdorur për ndërtimin e një modeli matematiko-statistik që analizon gjendjen shëndetesore të banorëve në një zonë industriale .

Konkretisht të dhënat janë marrë nga Instituti i Shendetit Publik, dhe përfshijnë informacione përdorur gjendjen e përgjithshme shëndetësore, për problemet ne pneumologji, të dhëna të fushes hematologjike , dermatologjike, etj.

Për ndërtimin e nje modeli të saktë statistikor , u shfrytëzuan edhe të dhëna të të tjera specifike për problemet e trajtuara.

1 Analiza e variancës një përmasore

1.1 Teoria bazë e ANOVA-s

ANOVA, (shkurtim i fjalëve Analizë e Variancës), është **iniciuar** rreth viteve 1920 nga statistikani anglez Sir Ronald Aylmer **Fisher**. Metoda synon të testoje hipotezën që disa variabla kanë të njëjtën mesatare në një apo më shumë popullime, apo jo.

Fisher bëri një varg arsyetimesh mbi variancen dhe arriti ne konkluzione te rëndësishme. Bazë e këtyre arsyetimeve ishte ndarja e variancës së zgjedhjes të një popullimi, në dy komponente dhe, krahasimi i këtyre komponenteve **duke përdorur statistikën Fisher**. Këto janë dhe arsyet pse ANOVA njihet ndryshe me emrin ANOVA e Fisherit ose analiza e Fisherit.
Pa dyshim zhvillimi i teknologjisë kompjuterike dhe rritja e fuqisë dhe shpejtësisë së tyre përpunuese, bënë që efektet e kësaj metode statistikore të jenë dramatike.

Le të shqyrtojme kete metodë statistikore.
Janë dhënë i nënpopullime të një popullimi. Shënohet μ_i mesatarja e nënpopullimit të *i*-të për i=1,2,...,I. Duke marrë zgjedhje të rastit të i nënpopullimeve, vleresohen pikërisht këto mesatare. Pastaj, kërkohet të kontrollohet hipoteza:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_i = \mu,$$

sipas së cilës, të gjitha nënpopullimet e popullimit kanë pritje matematike të njëjtë.
Kjo hipoteze ka *hipotezë alternative*, ka te pakten dy nenpopullime me pritje matematike te ndryshme.
Për të kontrolluar Ho, supozohet se jemi duke shqyrtuar nënpopullime normale që kanë variancë të njëjtë, pra që nënpopullimet kanë ligj $N(\mu_i, \sigma^2)$
Hipotezës së mësipërme mund ti jepet forma :

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_i = 0,$$

(efektet e nënpopullimeve janë 0)

Dihet se per popullimin ne shqyrtim, $MS_T = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \bar{y})^2/(n-1)$ (1) është vlerësim i pazhvendosur për σ^2. Pikërisht vlerësimin e mësipërm të variancës, Fisher e ndan në dy komponente.

1.2 Modeli në nënpopullime të përcaktuara

Në modelin e nënpopullimeve të përcaktuara, shqyrtohen i nënpopullime të popullimit, të fiksuara nga eksperimentuesi. Në secilin prej nënpopullimeve të fiksuara merret një zgjedhje e rastit me vëllim J_i. Vlera e *j*-te e zgjedhjes se nënpopullimit të *i*-të shënohet y_{ij}.

Modeli ndertohet si me poshte:

Së pari: Shënohet me μ_i mesatarja e nënpopullimit të *i*-të per *i* nënpopullimet e popullimit, (nënpopullime këto të fiksuara nga eksperimentuesi). Ndaj, me supozimin se nënpopullimi i *i*-të ka shpërndarje $N(\mu_i,\sigma^2)$,cdo y_{ij} vlere e zgjedhjes te marre, mund të paraqitet ne formën: $y_{ij}= \mu_i+ e_{ij,}$ ku e_{ij} kanë shpërndarje $N(0, \sigma^2)$.

Së dyti: μ_i mund të shkruhen si shumë e μ (pritje matematike të popullimit) me α_i që tregon efektin e nënpopullimit të *i*-të. Përfundimisht vlerat e marra të zgjedhjes shkruhen në formën: $y_{ij}= \mu+ \alpha_i+ e_{ij}$

Ky model linear i fituar quhet *modeli linear i ANOVA-s* e, meqë nënpopullimet janë paracaktuar nga eksperimentuesi, justifikohet edhe emërtimi i këtij modeli, si **model në nënpopullime të përcaktuara**[1].

Pikërisht, ky model linear i ANOVA-s përdoret për vlerësimin pikesor te parametrave μ dhe α_i . **Metoda e Katroreve më të Vegjël**, siguron në bazë të teoremës së Gaus –Markovit, një vlerësim pikesor të pazhvendosur dhe me dispersion minimal të parametrave μ dhe $\alpha_{i,}$, kur ato fitohen si kombinim linear i vlerave të vrojtuara y_{ij}.

Pra, ndërtohet $S(\mu, \alpha_i) = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \mu - \alpha_i)^2$, dhe kërkohet të gjenden vlerat μ dhe α_i që minimizojnë këtë shumë.

$$\frac{\partial S}{\partial \mu} = -\sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \mu - \alpha_i) = 0$$

$$\sum_{i=1}^{I}\sum_{j=1}^{j_i} y_{ij} = n\mu + \sum_i\sum_j \alpha_i = n\mu + \sum_i J_i\alpha_i$$

Atehere vlerësimi për μ (me supozimin se $\sum_i J_i\alpha_i$=0) është

$$\hat{\mu} = \frac{\sum_i\sum_j y_{ij}}{n}$$

Ndërsa për të vlerësuar $\alpha_{i,}$ llogarisim derivatet e pjesshme sipas parametrit α_i

$$\frac{\partial S}{\partial \alpha_i} = -\sum_j (y_{ij} - \mu - \alpha_i) = 0$$

$$\sum_j y_{ij} = J_i\mu + J_i\alpha_i$$

$$\hat{\alpha}_i = \frac{\sum_j y_{ij}}{J_i} - \mu = \overline{y_i} - \overline{y}$$

Pra vlerësimet që merren me MKV janë

$$\hat{\mu} = \overline{y} \quad (*)$$

$$\hat{\alpha}_i = \overline{y}_i - \overline{y}$$

Kështu $\hat{\mu} + \hat{\alpha}_i = \overline{y} + (\overline{y}_i - \overline{y}) = \overline{y_i}$

Por dihet nga teorema Gaus –Markovit se, jo vetëm vlerësimet që merren për parametrat janë të pazhvendosur, por dhe se këto vlerësime sigurojnë një minimum të

variancës nga të gjithë vlerësuesit e pazhvendosur të μ dhe $\alpha_{i,}$ që fitohen si kombinim linear i vlerave të vrojtuara.
Kështu, shuma e katrorëve të gabimeve

$$\text{SSB} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \overline{y_{ii}})^2$$

duke u pjestuar me numrin e gradëve të lirisë të saj, siguron një vlerësim të pazhvendosur minimal të σ^2.
Një vlerësim i pazhvendosur për σ^2, është dhe mesatarja e variancave të grupeve të zgjedhjes. Duke marrë

$$MS_B = \frac{\sum_i s_i^2}{I} = \sum_{i=1}^{I}\sum_{j=1}^{Ji}(y_{ij} - \bar{y}_i)^2 \Big/ I(J_i - 1) = \sum_{i=1}^{I}\sum_{j=1}^{Ji}(y_{ij} - \bar{y}_i)^2 \Big/ (n - I) \qquad (2)$$

nëse hipoteza H_0 është e vërtetë dhe, duke ditur për më tepër s_i^2 -të, kanë shpërndarje $\chi^2(j_i\text{-}1)$, MS_B do të ketë shpërndarje χ^2 $(I(j_i\text{-}1))$, pra $\chi^2(n\text{-}I)$.

Tani, shuma e dyfishtë (1) transformohet në shumën (2) që, nëse hipoteza eshte e vërtetë, siguron një tjetër vlerësim të pazhvendosur për σ^2.
Duke u nisur nga barazimi $y_{ij} - \bar{y} = (y_{ij} - \overline{y_i}) + (\overline{y_i} - \bar{y})$ dhe pasi ngrihen në katrorë dy anët e tij, shumohet sipas *i*-ve dhe *j*-ve dhe merret

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \overline{y_i})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{y}_i - \bar{y})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \overline{y_i})(\bar{y}_i - \bar{y})$$

Meqënëse termi i fundit i kësaj shume mund te transformohet në trajtën (2) $\sum_{i=1}^{I}(\bar{y}_i - \bar{y})\sum_{j=1}^{J_i}(y_{ij} - \bar{y}_i)$,duket qartë që ky praktikisht është 0. Ndaj, shuma merr formën

$$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \overline{y_i})^2 + \sum_{i=1}^{I} J_i(\bar{y}_i - \bar{y})^2 = SS_B + SS_M$$

Duke u nisur nga barazimi $SS_T=SS_B+SS_M$ dhe nga përfundimet (1) dhe (2), në bazë të Teoremës Koçran, SS_M dhe SS_B janë të pavarur dhe janë vlerësime për σ^2 , madje $SS_M/(I\text{-}1)$ ka shpërndarje $\chi^2(I\text{-}1)$.
Vërehet se SS_M eshte shuma e katroreve te gabimeve të mesatareve të grupeve me mesataren e tërë zgjedhjes.
Kështu, MS_B/ MS_M ka shpërndarje $F(I\text{-}1,n\text{-}I)$, si raport i dy statistikave të pavarura me shpërndarje χ^2, raport ky, bazë ne analizën ANOVA.
Në programet kompjuterike, kur japim komandën për vlerësim me metodën ANOVA, afishohen:
vlera g_0 e testit të mësipërm, pra vlera e raportit Fisher per zgjedhjen
p- vlera koresponduese e saj sipas tabelës së Fisherit (pra llogaritet probabiliteti që testi statistik të marrë vlerën g_0 ose vlera me ekstreme se g_0 ,kur H_0 është e vërtetë).
Gjeometrikisht, p-vlera, paraqet dyfishin e siperfaqes në të djathte të vlerës së vrojtuar, nën grafikun $F(I\text{-}1,n\text{-}I)$. Kështu mund të përcaktojmë jo vetëm se hipoteza

është apo jo e vërtetë, por dhe nivelin e gabimit të llojit të parë me të cilin ne mund ta pranojmë këtë hipotezë nëse është e mundur.

1.3 Modeli në nënpopullime të rastësishme

Tabela e ANOVA-s shfrytëzohet edhe në një këndvështrim tjetër.
Për këtë, më parë ndërtohet një tjetër model, që quhet **modeli në nënpopullime të rastësishme**[3].
Shihet popullimi N(μ,σ^2), dhe, lidhur me të, edhe tërë nënpopullimet e ketij popullimi të përcaktuar sipas një faktori. Me anë të zgjedhjes së marrë, ndërtohet një zgjedhje për popullimin e nënpopullimeve.
Për këtë, llogariten më parë mesataret m_i për cdo zgjedhje nga grupet e marrë sipas një faktori, dhe mendohen nënpopullime me shpërndarje $N(mi, \sigma^2)$, për i $=1,2,..I$, që tani shërbejnë si zgjedhje me vëllim i për popullimin e nënpopullimeve të popullimit te madh.
Me σ_a shënohet varianca e vlerave të kësaj zgjedhje nga μ.
Kështu, në bazë të modelit linear, elementet e kësaj zgjedhje mund të shkruhen në formën:
$m_i = \mu + a_i$, ku a_i e pavarura kanë shpërndarje normale me mesatare 0 dhe variancë $\sigma_a^{2(10)}$.
Përfundimisht, vlerave të zgjedhjes fillestare të popullimit i jipet një paraqitje tjetër në formën :
$y_{ij} = m_i + e_{ij} = \mu + a_i + e_{ij,}$ ku a_i e pavarura kanë shpërndarje normale me mesatare *0* dhe variancë σ_a , ndërsa $e_{ij,}$ gjithashtu të pavarura, kanë shpërndarje $N(0, \sigma^2)$.
Në këtë model të ndërtuar, interesohemi për σ_a^2 që paraqet variacionin e pritjes matematike të nënpopullimeve të përcaktuara sipas një faktori, nga pritja matematike e popullimit μ.
Shtrohet hipoteza

$$H_0: \sigma_a^2 = 0 ,$$

(domethënë që nuk ka variacion mes nënpopullimeve sipas faktorit në studim),
me hipotezë alternative

$$H_1: \sigma_a^2 \neq 0.$$

Duke përdorur Metodën e Katrorëve më të Vegjël, merret përsëri i njëjti vlerësim për pritjen matematike të popullimit, si në rastin e modelit në nënpopullime të përcaktuara,

$$\hat{\mu} = \sum \frac{m_i}{I} = \bar{y}$$

Nga teorema Kocran merret dhe një vlerësim i pazhvendosur për σ_a^2, që jipet nga shuma e mëposhtme $\sum_i (m_i - \hat{\mu})^2 = \sum_i (m_i - \bar{y})^2$, pjestuar me I-1.

Por, dihet se m_i fitohen nga y_{ij} për i të fiksuar dhe $j=1,..J$ sipas barazimit $y_{ij} = m_i + e_{ij}$, ku $e_{ij,}$ kanë shpërndarje $N(0, \sigma^2)$.

Ndaj SS_M / (I-1)= $\sum_{i=1}^{I} J(\bar{y}_i - \bar{y})^2$/(I-1) është një mesatare e vlerësimeve të variacionit të grupeve, pra përmban edhe variacionin brenda cdo grupi $\sigma^{2,}$ dhe J herë variacionin nga grupi në grup, σ_a^2, (me supozimin që vëllimi i zgjedhjeve në cdo grup është marrë i njëjtë), ndërkohë që dihet se vlera e kësaj statistike afishohet në tabelën ANOVA, dhe për më tepër se statistika ka shpërndarje χ^2(I-1)
Duke bërë të njëjtën ndarje të katrorit të gabime të popullimit, si në rastin e modelit të parë, kemi :

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{I} J_i(\bar{y}_i - \bar{y})^2 = SS_B + SS_M$$

SS_B/ (n-I) është një vlerësim i pazhvendosur për variacionin brenda grupeve që është σ^2, me shpërndarje të njohur χ^2(n-I) (mund të shihet si mesatare e variacionit brenda cdo grupi, që e shënuam σ^2).
Për të vlerësuar σ_a^2, bazohemi në vlerësimin:
$\hat{\sigma}_a^2$ =(MS_B-MS_M)/J (ku J është vëllimi i njëjtë i zgjedhjes i marrë në cdo grup)
Kështu për testimin e hipotezës H_0: $\sigma_a^2 = 0$ do të shfrytëzojmë po tabelën ANOVA dhe rezultatin e afishuar të statistikës së Fisherit.

1.4 Kritere dhe vërejtje në aplikimin e ANOVA-s

1. Metoda ANOVA kerkon, para aplikimit te saj, te kontrollohen nëse plotësohen apo jo dy kushtet e sipërpërmendura, normalizimi i të dhënave si dhe kontrolli i barazimit të variances mes nënpopullimeve në studim.
Metodat e perdorura për kontollin e normalizimit të të dhënave është **metoda Kolmogorov-Smirnov**-it dhe **Shapiro-Wilk**-it .
Ndërsa për të kontrolluar kushtin e dytë të barazimit të variancave përdoret **testi Levene-**t, në format e tij të përpunuara (shih Testin Levene, 1.5).
2. Metoda e ANOVA-s përdoret shume vecanërisht kur sipas niveleve të faktorit në shqyrtim, bëhet ndarja e popullimit të madh në disa (më shumë se dy) nënpopullime.
Nëse ndarja do të ishte në dy nënpopullime atehere trajtimi i problemit si me metoden e ANOVA-s do të jipte te njëjtin rezultat sikur ai të trajtohet me ane të t testit.
3. Lind pyetja në këto raste cila nga metodat është me e pershtatshme. Në kushtet e mosplotesimit te Kushtit të Homogjenitetit të Variances mes grupeve, metoda me mire është ajo e t-testit, sepse me modifikimin e bere ajo lejon përdorimin e saj në dy raste dhe kur plotësohet dhe kur nuk plotësohet kushti në fjalë.
Kur Kushti i Homogjenitetit të Variancës plotësohet përdoret t-testi Student, ndërsa kur ky kusht nuk plotësohet përdoret t-testi Welch,
Ndërsa në kushtet e më shumë se dy nënpopullimeve, përdorimi i t-testit për cdo dyshe të studiuar s'do të kishte kuptim sepse për cdo t-test vendimi do të merret me një gabim të llojit të parë α, kështu do të rritej shumë gabimi në marrjen e vendimit që te gjitha nënpopullimet janë statistikisht me mesatare të njejte.
4. F testi për hipotezën $\sigma_a^2 = 0$ në modelin e dytë te metodës në fjalë është një vlere e përafruar e testit në fjalë, kur J_i s'janë te barabarata. Megjithate kjo pranohet si procedurë praktike, derisa testi i sakte do të ishte shumë i komplikuar.

1.5 Testi Levene (Metoda e Homogjenitetit të Variancave)

1.5.1 Vështrim i përgjithshëm i testit

Testi Levene përdoret në statistike për të kontrolluar barazimin apo jo te variancës mes dy ose më shumë grupeve.
Shume teste statistikore e kanë si kusht supozimin se dispersionet e popullimeve nga te cilat janë marrë zgjedhje të ndyshme, janë të barabarta. Procedurat tipike ku ky kusht kërkohet është Analiza e Variancës (ANOVA) dhe t- testi.
Testi i Levene përdoret zakonisht para krahasimit te mesatareve. Duke u nisur nga rezultatet e ketij testi vendoset menyra që do të përdoret për krahasimin e mesatareve, do të jetë ajo parametrike apo jo parametrike, metode kjo e fundit që është e lire nga kushti i barazimit të variancave.
Ai mund të perdoret edhe si test, thjesht për të përcaktuar se dy ose me shume zgjedhje të marra nga i njëjti popullim kanë apo jo variancë të barabartë.
Testi njihet ndryshe me emrin Testi i Homogjenitetit të Variancës.
Testi kontrollon hipotezën H_0 që variancat janë të barabarta, me hipotezë alternative jo të gjitha variancat (dispersionet e popullimeve) janë të barabarta.
Nëqoftëse p-vlera e rezultatit të këtij testi është e vogël, më e vogël se një nivel rëndësie, (zakonisht 0.05), atëhere diferencat në variancat e zgjedhjeve kanë dhënë rezultate të tilla që nuk ndodhin nëse zgjedhjet e rastit do të merreshin nga popullime me variancë të njëjtë.
Përfundimisht, hipoteza, që variancat janë të barabarta, bie poshtë dhe arrihet në përfundim që ka diferencë mes variancave (dispersioneve) të popullimeve.

1.5.2 Përcaktimi i testit

Testi statistik W përcaktohet si më poshtë:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2},$$

ku *W* është rezultati i testit
k numri i grupeve që i nënshtrohen testit,
N numri total i rasteve ne gjithë grupet,
N_i numri i rasteve në grupin e i-te
$Zij=|Yij-\bar{Y}i.|$ ku Y_{ij} eshtë vlera e variablit të matur për elementin e j-te në grupin e *i*-të ndërsa $\bar{Y}i.$ mesatarja e grupit të *i*-të.
ose $Zij=|Yij-\tilde{Y}i.|$ ku $\tilde{Y}i$ është kuartili i dyte i grupit
Të dy vlerësimet përdoren.
Sipas vlerësimit të pare, shmangia e vlerave të vrojtuara behet duke llogaritur distancën e tyre nga mesatarja e grupit. Mesataret e grupeve konsiderohen si vlerësime pikësore për mesataren e popullimit, të cilit zgjedhja i përket.Vlerësimi në fjalë përdoret kur të dhënat e zgjedhjes plotësojnë kushtin e normalizimit.
Ndërsa në formën e dytë, si bazë në llogaritje merren distancat nga kuartilet e dyta, median-at e vlerave te vrojtuara të zgjedhjes të popullimit në studim. Metoda e dytë njihet si **testi Brown-Forsythe**. Ajo përdoret, dhe jep rezultate te mira, në rastet kur të dhënat e zgjedhjes nuk plotësojnë kushtin e normalizimit.
$Z_{\cdot\cdot}$ është mesatarja e gjithë Z_{ij} pra

$$Z_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} Z_{ij}$$

$Z_{i.}$ është mesatarja e grupit të *i*-te

$$Z_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} Z_{ij}$$

Testi i përcaktuar W provohet se ka shperndarje Fisher me parametra(k-1, N-k).
Ndaj për marrjen e vendimit krahasohet vlera e vrojtuar me vleren kritike te testit, që përcaktohet nga F(k-1, N-k) për një nivel besimi të zgjedhur (0.05 ose 0.01).
Në përpunimin kompjuterik të të dhënave procesi i marrjes së vendimit bëhet sic është sqaruar në paragrafin e mësipërm.
Testi e Levene konsiderohet robust dhe i fuqishem. Qëndrueshmëria e testit lidhet me aftësinë e testit për të arritur në konkluzionin që variancat s'janë të barabarta kur në fakt ato s'janë të tilla. Robuste lidhet me faktin qe testi arrin ne konkluzionin e gabuar që variancat s'jane të barabarta, kur zgjedhja s'ka shpërndarje normale dhe në të vërtetë ato janë të barabarta.
Punimi origjinal i Levene në fakt propozohet vetëm me mesataren. Ishte Brown dhe Fosythe që më 1974 e përmirësuan testin ne fjale dhe perdoren medianen në këtë vlerësim.
Me ane te studimeve të kryera me metodën Monte Karlo, u tregua se perdorimi i mesatares vlen nese te dhenat tregojnë për shpërndarje simetrike dhe plotësojnë kushtin e normalitetit. Ndërsa, kur të dhënat nuk plotësojnë kushtin e normalitetit, psh kur kemi shperndarje Koshi, rezultet më të mira i siguron përdorimi i medianës.

1.6 Krahasimi i shumëfishtë

1.6.1 Një trajtim i përgjithshëm i problemit

Në problemin e trajtuar lind pyetja: Cili nga ciftet e mesatareve të nënpopullimeve është i ndryshëm?
Një variant është të analizojmë të gjithë ciftet e mundshme të nënpopullimeve duke krahasuar pritjet matematike. Në rastin tonë kemi për të kontrolluar disa hipoteza H_0:$\mu_i=\mu_j$, *për* $i\neq j$
Lind problemi i studimit të nënpopullimeve që i korespondojnë nivele të ndryshme të këtij faktori.
Për të përcaktuar se cili nga nivelet do të konsiderohet i ndryshëm, duhen dalluar mesataret statistikisht të ndryshme nga njëra-tjetra.
Matematikisht kemi të bëjmë me problemin e **krahasimit te shumëfishtë**, të shumëfishtë në kuptimin se kemi të bëjmë me disa krahasime të njëpasnjëshme. Për cdo krahasim te tille, pergjigjja e studimit të një hipoteze, jepet me nje koeficient besimi te lartë. Pra, mundesia qe kjo hipotezë të anullohet, ndërsa ajo është e vërtetë, është shumë e vogel. Por, nese bëhen shumë krahasime të njëpasnjëshme, mundësia që të gjitha këto hipoteza të hidhen poshtë, ndërsa ato janë të vërteta, rritet ndjeshëm.

Shume matematikanë kanë paraqitur teknika të ndryshme për kontrollimin e këtij gabimi, në rastin e krahasimeve të shumëfishta. Teknikat që ato propozojnë, tentojnë një reduktim të gabimit të llojit të parë për secilin krahasim, në mënyrë të tillë që megjithëse gabimi i shumë krahasimeve të rrisë α-ën, ajo te mos kalojë vlerën e dëshiruar.
Por ka dhe metoda që synojnë rritjen e zonës të vlerave të lejuara për diferencën midis cdo dy mesatareve qe vlerësohen, në mënyrë që, ndërprerja e këtyre zonave pas të gjitha krahasimeve të nevojshme, të japë një gabim të llojit të parë , sa ai i dëshiruar.
Një metodë e tillë është **metoda Tukey**, e përdorur në rastin kur për vlerësimin e nënpopullimeve përdoren zgjedhje nga nënpopullime që plotësojnë kushtin e barazimit të variancës, dhe **Scheffe**, që përdoret në rastin kur kushti në fjalë nuk plotësohet.

1.6.2 Metoda Tukey

Kjo metode ka marrë emrin e matematikanit dhe kimistit amerikan të shekullit XX, John Tukey[3].
Ajo synon të dallojë mesataret e ndryshme nga njëra-tjetra duke i krahasuar ato dy e nga dy.
Le ta shohim metoden Tukey, duke trajtuar disa cështje të saj, supozimet ku bazohet kjo metode, testin që ndërton për vlerësim, statistikën në të cilën bazohet si dhe rradhën e krahasimit që propozon
Metoda Tukey ka në bazë të saj *disa supozime*. Së pari, vëzhgimet janë të pavarura dhe, së dyti, mesataret, që studiohen, janë nga popullime me shpërndarje normale. Por kushti qe e dallon perdorimin e metodes Tukey nga metodat e tjera është që vëllimi i zgjedhjeve të nënpopullime është i njejtë.
Testi Tukey bazohet në testin e zakonshëm Studenti, ose Gosset, qe përdoret për krahasimin e dy mesatareve.
Shembull i thjeshtë:
Duam të kontrollojmë hipotezën H_0: $\mu_1=\mu_2$
me hipotezë alternative që pritjet matematike të dy popullimeve, janë të ndryshme .
Ky problem është i njëjtë me H_0: $\mu_1-\mu_2=0$, me H_1: $\mu_1-\mu_2 \neq 0$

Për vlerësimin e kësaj hipoteze sipas metodës Studenti , përdoret statistika $T=\frac{\bar{x}-\bar{y}}{\frac{s}{\sqrt{n}}}$

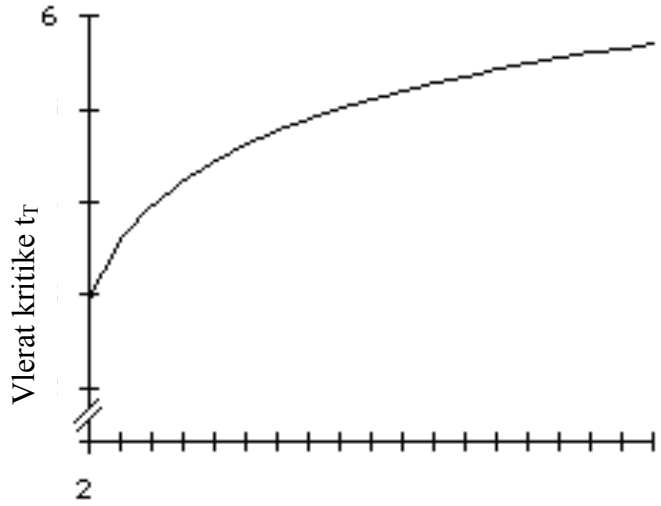
që ka shpërndarje Studenti me 2n-2 shkallë lirie (n është vëllimi i zgjedhjeve që do të shqyrtohen dhe, ku në këtë formulë është supozuar të jetë i njëjtë. Në rast se ky vëllim nuk është i njëjtë, atëhere në vend të *n* në formulë do të punohet me mesataren e dy vëllimeve të zgjedhjeve). Pasi përcaktohet zona e vlerave të lejuara (-t, t), duke shfrytëzuar tabelën e shpërndarjes Studenti për α-ën, dhe 2*n-2* shkallë lirie, shihet nëse vlera e testit të marrë bie në zonën kritike. Në këtë rast themi se popullimet kanë mesatare të ndryshme.
Por, përdorimi i këtij testi në tërë krahasimet e cifteve të popullimeve, thamë, se do të sjellë rritjen e gabimit të llojit të parë.
Testi Tukey mundohet të zgjidhë këtë problem, duke modifikuar zonën e vlerave të lejuara të statistikës Studenti, në varësi të numrit të mesatareve që do të kontrollohen. Kuptohet se sa më i madh të jetë ky numër, aq më e madhe duhet marrë vlera kritike e

modifikuar $t_{T,}$ duke berë që dhe zona e diferencës së lejuar të dy mesatareve të një krahasimi të vetëm të jetë më e madhe.
Për të kuptuar më mirë këtë ndryshim midis zonës së vlerave të lejuar të marra sipas metodës Studenti, dhe asaj Tukey, le ti kthehemi problemit të thjeshtë të mësipërm, problemit të krahasimit vetëm të dy mesatareve.
Per kontrollin e kesaj hipoteze, sipas Studentit ndërtohet statistika:

$$T = \frac{\bar{x} - \bar{y}}{\frac{s}{\sqrt{n}}}$$

Kurse, sipas Tukey, ndërtohet statistika $T_{T=} \frac{\bar{x} - \bar{y}}{\frac{s\sqrt{2}}{\sqrt{n}}}$, ku 2-shi përcaktohet nga numri i mesatareve që po krahasohen.
Kuptohet që ky ndryshim në test do të rrisë zonën e vlerave të lejuar nga (-t,t) sipas metodës Studenti, në $(-t\sqrt{2}, t\sqrt{2})$ sipas metodës Tukey.
P.sh, nëqoftëse. në cdo grup kemi marrë një zgjedhje me vëllim 13, (n=13), atëhere për dy grupet që krahasohen, do të arrihej në përfundim që kanë pritje matematike të ndryshme, nëse vlera e testit T është në $(-\infty,-2.06)$ ose në $(+2.06,+\infty)$. Por me modifikimin e bërë sipas Tukey, ky test do t'i takonte zonës kritike, pra nënpopullimet do të konsiderohen me pritje matematike të ndryshme, nëse vlera e testit në fjalë është në $(-\infty,-2.06*1.414) = (-\infty,-2.91)$ose në $(+2.91,+\infty)$. Dhe nga tabela Studenti, kësaj vlere kritike i korespondon $\alpha = 0.025$.
Nga dy krahasimet që janë të nevojshme të bëhen në këtë rast, probabiliteti që të dy nënpopullimet të konsiderohen me pritje matematike të njëjtë, është $(1- \alpha)^2=(1-0.025)^2 = 0.95$, duke siguruar kështu përfundimisht një përgjigje me nivel besimi 0.05, sa ai i kërkuari.
Duhet theksuar se, sipas kësaj metode, në rast se vëllimi i cdo zgjedhjeje të marrë që do të krahasohet është i njëjtë , koeficientin e besimit, e ka pikërisht $1- \alpha$.
Nga ky sqarim shihet se, kur numri i mesatareve që do të kontrollojmë është 2, $t_T= t\sqrt{2}$, pra testi Tukey dhe Studenti janë identik.
Problemi i ndryshimit midis dy testeve do të shfaqet nëse krahasojmë më shumë mesatare.
Në tabela të vecanta të paraqitura, mund të lexohen vlerat e t_T për një nivel besimi të dhënë α dhe, për një numër të përcaktuar të zgjedhjes së nënpopullimeve n, që marrin pjesë në këtë krahasim të shumëfishtë. Në grafikun e paraqitur mund të lexohen këto vlera t_T per $\alpha = 0.05$ ne varësi të numrit të mesatareve që do të krahasojmë.

Figura 1–1 Vlerat Kritike – numri i mesatareve qe krahasen



Tukey kujdeset edhe për reduktimin e numrit të krahasimeve të mesatareve të popullimeve. Psh, në rastin e shqyrtimit të mesatareve (A,B,C,D), të cilat supozojmë se i kemi renditur sipas rendit A>B>C>D, sipas testit Tukey, nuk është patjetër e domosdoshme të bëhen të gjitha krahasimet e mundshme.

Për këtë sugjerohet kjo *radhë krahasimi*: Krahasohet më parë popullimi me mesataren më të madhe (A), me atë më të vogël (D). Nëqoftëse t_T (vlera e statistikës që Tukey propozon për vlerësim) është më e vogël se vlera t_T e shpërndarjes së propozuar, hipoteza Ho pranohet si e vërtetë. Pra, mesataret e këtyre zgjedhjeve nuk janë statistikisht të ndryshme. Kuptohet, derisa nuk kanë ndryshim këto dy mesatare që kishin diferencën më të madhe, krahasimi i mesatareve të popullimeve të tjera është i panevojshem. Procesi i krahasimit të çifteve ndërpritet.

Përfundimisht, për rradhën e krahasimit në testin Tukey, fillohet gjithnjë me krahasimin e mesatares më të madhe me atë më të vogël. Pastaj vazhdohet me krahasimin e kësaj mesatareje më të madhe , me mesataren e dytë më të ulët, derisa mesatarja më ë lartë të jetë krahasuar me të gjithë mesataret e tjera, ose sic thamë, derisa të mos ketë ndryshim mes mesatareve në studim. Pas ketij krahasimi, i njëjti proces përsëritet me mesataren e dytë për nga madhësia: krahasohet me mesataren më të vogël, pastaj me atë të dytë më të ulët, e kështu me rradhë.

Procesi i krahasimit përfundon ose kur janë bërë të gjithë krahasimet e mundshme e ka ndryshim mes cdo cifti që krahasohet, ose sa gjendet cifti i pare i zgjedhjeve mesataret e të cilëve s'kanë ndryshim statistikor .

Si konkluzion, mund të themi se metoda Tukey është një metodë e krahasimit të shumëfishtë që ka një kontoll të dobët të nivelit të besimit, në kuptimin që në cdo krahasim ciftesh ajo nuk kontrollon drejtperdrejt nivelin e besimit për këtë krahasim, por ndërhyn në mënyrë të tërthortë në të.

Me këtë metodë, në rast se vellimi i cdo zgjedhje të marrë që do të krahasohet, është i njëjtë, koeficienti i besimit është pikërisht 1- α Në programet kompjuterike statistikore mund të përdoret lehtësisht kjo metodë e krahasimit te shumëfishtë

(Tukey). Kjo do të na afishojë gjithë intervalet e besimit të hipotezave $\mu_i = \mu_j$, ose $\mu_i-\mu_j=0$, me koeficient besimi 0.95. Kështu, thjeshtë mund të identifikohet cili nga ciftet e nënpopullimeve, ka mesatare të ndryshme. Dy nënpopullime konsiderohen me mesatare të ndryshme nëse intervali i tyre i besimit nuk përmban 0.

1.6.3 Metoda Scheffe

Metoda Scheffe[1] është metodë e krahasimit të shumëfishtë, por ky krahasim bëhet në një hap të vetëm. Kjo përben dhe thelbin e ndryshimit të saj me metodën e mësipërme Tukey, e cila duke krahasuar dyshe nënpopullimesh, zhvillohej në disa etapa.
Kjo është një ndër metodat bazë që përdoret në analizën e regresit linear për të realizuar krahasimin e shumëfishtë, dhe natyrisht dhe në ANOVA-n (si një rast i vecantë i regresit linear) Metoda ka marre emrin e statistikanit amerikan Henry Scheffe.
Ndërtohet një kontras C i përcaktuar nga barazimi

$$C = \sum_{i=1}^{r} c_i \ \mu_i$$

dhe shtrohet hipoteza H_0: C = O me hipotezë alternative $C \neq 0$
Vlerësimi i C behet sipas barazimit te mëposhtëm

$$\hat{C} = \sum_{i=1}^{r} c_i \bar{Y}_i$$

dhe varianca e së cilës është

$$s^2_{\hat{C}} = \hat{\sigma}^2_e \sum_{i=1}^{r} \frac{c_i^2}{n_i},$$

ku n_i është vëllimi i zgjedhjes së marrë nga nënpopullimi i *i*-të , dhe
$\hat{\sigma}^2_e$ varianca e vlerësuar (estimated)
Forma e intervaleve të besimit, me koefficient besimi $1-\alpha$, është:

$$\hat{C} \pm \ s_{\hat{C}} \sqrt{(r-1)\, F_{\alpha;r-1;N-r}}$$

ku N është vellimi i zgjedhjes së gjithë popullimit, r numri i nënpopullimeve në të cilat është ndarë popullimi i madh.
Pas afishimit të ketij intervali, kontrollohet nëse ai e përmban apo jo zeron. Nëse nuk e përmban, atëhere hipoteza H_0 bie.
Metoda në fjalë konsiderohet shume rigoroze, sepse kontrollet i ben per një gabim të rendit shumë më të vogël se α.

1.6.4 Krahasimi i dy metodave

Metoda Tukey përdoret në nenpopullime me vëllime të njëjta, ndërsa metoda Scheffe nuk vendos nje kufizim të tille.
Metoda Scheffe nuk e ka të domosdoshëm kushtin që $\sum_1^n c_i = 0$ por, nëse ky kusht plotësohet ne vendosjen e konstanteve, atëherë themi se po përdoren koeficientë qe

plotesojnë **kushtin e kontrastit**. Ndërsa metoda Tukey e ka domosdoshmeri qe gjate krahasimit të cifteve të plotësohet kushti i kontrastit.
Së fundmi, metodat në fjalë bazohen në statistika me shpërndarje të ndryshme. Metoda Tukey bazohet ne statistiken me shperndarje Studenti, ndersa ajo Scheffe ne shperndarjen Fisher. Është e kuptueshme që rezultatet e dy metodave në fjalë të jenë të ndryshme.
Përfundimisht, në rastin e krahasimit të mesatareve të tri nënpopullimeve me vëllim të njëjtë zgjedhje, interval më të ngushtë jep metoda Tukey, ndërsa ne rastet e 4 e me shumë nënpopullimeve interval më të ngushtë e siguron metoda Scheffe (bazuar kjo nga punimi i O'Neil dhe Wetherill më 1971)[3]
Metoda Scheffe coi në trajtimin e një problemi tjetër, atij të Analizes së Kontrastit[2]. Është pikërisht kjo analizë ajo që përcakton tërsisht modelin e ANOVA-s me anë e vlerave të këtyre koeficientëve.
Kontrastin si të thuash mes nënpopullimeve ajo e përkthen në bashkësi numrash, që janë pikërisht koeficientët e kontrastit (cij quhen edhe koeficientë të kontrastit).
Në varësi të rastit që shqyrtohet, koeficientët e kontrastit mund të jenë të planifikuar ose jo. Të përcaktuar janë në rastet kur eksperimenti kryhet për të percaktuar një hipoteze me koeficientë të paracaktuar. Pra, cdo gjë është mirë e planifikuar. Ndërsa rasti tjetër është kur këto koeficientë varen nga rezultatet e vrojtimit qe është kryer. Metoda e Kontrastit, në këtë rast quhet Post Hoc.
Nëse koeficientët nuk e plotësojnë kushtin e kontrastit, atëhere këto koeficientët tregojnë koeficientin e pjesshëm te korelacionit mes variablit në shqyrtim e variablit të varur. Pra percaktojne kontributin e variablit ne shqyrtim në variablin e varur.

1.6.4.1 Zbatim 1

Një perdorim i metodes Scheffe me koeficiente qe nuk plotesojne kushtin e kontrastit, është bërë në trajtimin e **problemit të ndikimit te SO_2 tek prodhueshmëria e bimëve** (shih problemin e trajtuar tek Kapitulli 6).
Në rastin në fjalë, pas aplikimit të metodës Leven, në krahasimin e variancës se nënpopullimeve tek tri komponentët (patate, perime, fasule) u morën vlera signifigative (shih Tabelen 1-1). Pra, në bazë të zgjedhjes së marre, u tregua per hedhje poshtë të hipotezës së barazimit të variances, duke na penguar në përdorimin e ANOVA-s .
Atëhere u përdor metoda e t-testit Welch,që tregoi ndikim negativ të zonës së ndotur, pra që tri komponentët e vektorit bimë kanë prodhueshmeri me të larte tek zona e kontrollit se tek zona industriale, e ndotur. U shtrua problemi i përcaktimit se në c'shkallë ishte ky ndikim.

Tabela 1–1 Afishimet e Testit Levene

	Levene's Test for Equality of Variances	
	F	Sig.

grure	Equal variances assumed	.234	.629
miser	Equal variances assumed	2.447	.120
patate	Equal variances assumed	31.638	.000
perime	Equal variances assumed	21.773	.000
fasule	Equal variances assumed	4.941	.028
jonxhe	Equal variances assumed	16.393	.000
fruta	Equal variances assumed	1.313	.254

Konkluzioni i lartpërmendur (1.6.4 Krahasimi i dy metodave) tregoi se metoda e Kontrastit mund te perdoret edhe ne krahasimin e dy nenpopullimeve me ane te t-testit. Koeficientët e Kontrastit që nuk plotësojnë kushtin e kontrastit, përcaktojnë tëresisht koeficientet [2] me te cilën lidhen mesataret e dy nënpopullimeve.
Pas shkrimit te kodit ne SPSS:

ONEWAY patate perime fasule BY area
/CONTRAST=1 -1.3
/MISSING ANALYSIS.

Përdorimi i Metodes së Kontrastit gjatë t-testit, dha afishimet e mëposhtme (Tabela 1-2):

Tabela 1–2 Rezultatet e Metodës së Kontrastit

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
patate	Assume equal variances	1	67.8788[a]	17.45542	3.889	153	.000
	Does not assume equal variances	1	67.8788[a]	23.76892	2.856	47.916	.006
perime	Assume equal variances	1	26.1727[a]	14.10142	1.856	153	.065
	Does not assume equal variances	1	26.1727[a]	17.38772	1.505	58.402	.138
fasule	Assume equal variances	1	12.4687[a]	5.55062	2.246	153	.026
	Does not assume equal variances	1	12.4687[a]	6.30780	1.977	71.498	.052

a. The sum of the contrast coefficients is not zero.

Rezultatet e afishuara treguan se prodhimi ne zonen e kontrollit te tri bimëve në fjalë është 130% e prodhimit të tek zonës së ndotur.

2 Modeli i regresit linear të shumëfishtë

2.1 Qëllimi i regresit të shumëfishtë

Qëllimi i metodës së regresit të shumëfishtë është analiza e lidhjes mes variablave te pavarur dhe nje variabli te varur. Variablat e varur mund të jenë metrik, pra ordinal apo i vazhdueshëm, apo dischotomous, pra variabla nominal me dy vlera. Variabli i varur duhet të jetë metrik.
Metoda e regresit kontrollon nëse një lidhje e tillë ekziston, e nëqoftëse ajo ekziston, synon të përdoret informacioni ekzistues për variablat e pavarur, që të përmirësohet saktësia në parashikimin e vlerës së variablit të varur.

2.2 Tipet e ndryshme të regresit te shumëfishtë

Njihen tri tipe të regresit te shumëfishtë, secila e dezinjuar në SPSS për tu përgjigjur pyetjeve te ndryshme:
Regresi i shumëfishtë standart që përdoret për të vlerësuar lidhjen mes nje bashkësie variablash te pavarura dhe një variabli të varur.
Regresi statistic (ose stepwise) ose që përdoret për të gjetur nënbashkësinë e variablave të pavarur (Stepwise) që kanë lidhje më të fortë me variablin e varur.
Regresi hirarkik që përdoret për të kontrolluar lidhjen mes një bashkësie variablash të pavarur dhe një variabli të varur, pasi më parë është kontrolluar për efektin e disa variablave të pavarur në variablin e varur.

2.3 Trajtimi matematik i regresit linear te zakonshëm, standart(GLM)

Le të paraqesim trajtimin matematik të këtij problemi të regresit linear të shumëfishtë. Supozojmë se variabli y varet nga p variabla "jo statistikisht të pavarur" $x_1,x_2,\ldots x_p$, që quhen variabla të pavarur. Problemi i regresit të shumëfishtë ka të bëjë me ndertimin e një modeli, ku y shprehet si funksion i këtyre variablave në formën

$$y= f(x_1,. x_2,\ldots,x_p) + \varepsilon \text{ , ku } \epsilon \text{ ka shpërndarje } N(0,\sigma^2),$$

Ndërsa, kur modelin e kërkojmë në formën
$y= \beta_{0+} \beta_1x_1 +\beta_2 x_2+...+\beta_px_p+\epsilon$, ku ϵ ka shpërndarje $N(0,\sigma^2)$, matematikisht kemi të bëjmë me modelin polinomial të regresit të shumëfishtë.
Në ndërtimin e këtij modeli duhen parë këto probleme:

2.3.1 Vlerësimi i parametrave të regresit linear të shumëfishtë (Hapi I)

Përcaktimi i zgjedhjes në shqyrtim mund të bëhet në dy mënyra :
Së pari, për një vlerë të fiksuar të variablave $x_1,x_2,...x_p$ bëjmë disa matje dhe marrim vlera të ndryshme të y_i, e më pas, fiksojmë një tjetër p-she vlerash të $x_1,x_2,...x_p$ e masim disa vlera të Y, duke marrë një zgjedhje të nënpopullimit në fjalë. Këtë proces

e vazhdojmë derisa të kemi marrë n vëzhgime. Sipas kësaj mënyre, vetëm Y shihet si variabël i rastit.
Së dyti, marrim n individë nga popullimi në fjalë dhe shikojmë, përcaktojmë, p+1 variablat për të cilën jemi të interesuar.
Kështu për të ndërtuar modelin e formës

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \varepsilon_i \qquad \text{për } i = 1..n$$

ne duhet të vlerësojmë $\beta_0, \beta_1, \beta_2, \ldots \beta_p$ që janë parametrat e këtij modeli dhe $\epsilon_1, \epsilon_2 \ldots \epsilon_n$, n variabla të pavarur $N(0,\sigma^2)$.
Modelin matematik që kërkojmë të ndërtojmë, mund ta paraqesim edhe në formë vektoriale[4].
Le të shënojmë me $\boldsymbol{\beta} = (\boldsymbol{\beta_0}, \ldots \boldsymbol{\beta_p})'$ vektorin parametrik $p + 1$ dhe $\boldsymbol{y} = (\boldsymbol{Y_0}, \ldots, \boldsymbol{Y_n})'$ vektorin i vëzhgimeve, ndërsa $e = (e_0, \cdots, e_n)'$ vektori i gabimeve dhe matrica e modelit $n * (p + 1)$

$$\begin{bmatrix} 1 & x_{11} & . & x_{1n} \\ 1 & x_{21} & . & x_{2n} \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{n1} & . & x_{pn} \end{bmatrix}$$

Kështu modeli matematik ka formën e këtij ekuacioni vektorial:

$$Y = X'\beta + e$$

ku e është një matricë $N(\mathbf{0}, \boldsymbol{\sigma^2 I})$

Për vlerësimin e parametrave $\beta_0, \beta_1, \beta_2, \ldots \beta_p$, përdorim metodën e katrorëve më të vegjël, pra ndërtojmë shumën e mëposhtme:

$$S(\beta) = \sum_i (y_i - \beta_o - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p)^2$$

Duke përdorur shënimet e mësipërme, matrica ku bazohet vleresimi mund të paraqitet në formë matricore $\boldsymbol{S} = (\boldsymbol{y} - \boldsymbol{X'\beta})'(\boldsymbol{y} - \boldsymbol{X'\beta})$.

$$\frac{\partial S}{\partial \beta} = 0 = -X(Y - X^T\beta) + (Y - X^T\beta)^T(-x^T)$$

$$0 = -XY + XX^T\beta - Y^TX^T + (X^T\beta)^TX^T$$

$$XX^T\beta = XY$$

Kështu vektori i koefiçientave të regresit $\boldsymbol{b} = (b_0, b_1, \cdots, b_p)'$ mund të llogaritet duke zgjidhur ekuacionin normal: $(XX')\beta = Xy$
dhe përcaktojmë vlerat e parametrave β duke kërkuar minimizimin e kësaj shume.
Zgjidhja e marë është

$$\boldsymbol{b} = (XX')^{-1}(Xy)$$

Ky vlerësim i parametrave β është një vlerësim i pazhvendosur , sepse:

$$\mathrm{E}(\hat{\beta}) = E(b) = E((XX^T)^{-1}(XY)) = (XX^T)^{-1}E(XY) = (XX^T)^{-1}XE(Y)_{=}$$
$$= (XX^T)^{-1}XX^T\beta = \beta$$

Nga teorema e Gaus-Markovit dimë, se vlera e parashikuar $\tilde{y}$ ka variancën minimale, nga të gjithë vlerësuesit e tjerë linearë të vlerave të X
$S = \sum_i (y_i - \tilde{y})^2$ na sherben si një vlerësim për σ^2. Madje, një vlerësim të pazhvendosur të σ^2, na e siguron dhe $MS_R = \sum_i (y_i - \tilde{y})^2/(n-p-1)$ ose ne trajte vektoriale vleresimi i pazhvendosur për variancën është $s^2 = (\boldsymbol{y} - \boldsymbol{X'}\,\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X'}\,\boldsymbol{b})/(\boldsymbol{n} - \boldsymbol{p} - \boldsymbol{1})$
dhe matrica e kovariancës është $cov(b) = \mathrm{s}^2(XX')^{-1}$
Paraqitet edhe një formë tjetër e *modelit të regresit linear të shumëfishtë*, forma e *qendërzuar*. Modeli i qendërzuar jepet nga:

$$\mathrm{y_i} = \beta_0' + \beta_1(\mathrm{x_{1i}} - \bar{\mathrm{x}}_1) + \cdots \beta_\mathrm{p}(\mathrm{x_{pi}} - \bar{\mathrm{x}}_\mathrm{p}) + \mathrm{e_i}, \qquad 1, \cdots, \mathrm{n},$$

ku

$$\bar{x_j} = \frac{1}{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{x_{jk}}\ , \qquad \mathrm{j} = 1, \cdots, \mathrm{p}, \quad \text{dhe} \quad \beta_0' = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_\mathrm{p}\bar{x}_\mathrm{p}$$

Vlerësuesit me Metoden e Katrorëve më të Vegjël të $\beta_1, \cdots, \beta_\mathrm{p}$, janë të njëjta si më sipër dhe emërohen $\mathrm{b}_1, \cdots, \mathrm{b}_\mathrm{p}$, ndërkohë që vlerësuesi me këtë metodë i β_0' është $b_0' = \bar{y}$. Avantazhi i këtij modeli është se vlerësuesit me Metoden e Katrorëve më të Vegjël $\mathrm{b}_1, \cdots, \mathrm{b}_\mathrm{p}$ janë të pakorreluar me b_0'. Kjo do të thotë që është më e lehtë të gjesh intervale besimi, duke u bazuar ne vlerat e parashikuara $\hat{y} = \bar{y} + b_1(x_1 - \bar{x_1}) + \cdots + b_p(x_p - \bar{x_p})$, siç do të shohim më poshtë.
Modeli i qendërzuar mund të shkruhet në formë matricore si më poshtë.
Le të jetë **A** matrica e shumes të katrorëve $p\,x\,p$ dhe prodhimit të devijimeve me elementët i, j $\quad a_{ij} = \sum_{k=1}^{n}(x_{ik} - \bar{x_i})\,(x_{jk} - \bar{x_j}), i, j = 1, \cdots, p.$ (*)
Le të jetë **g** vektori $p\,x\,1$ me elementin e i-të $g_i = \sum_{k=1}^{n}(y_k - \bar{y})\,(x_{ik} - \bar{x_i}), i = 1, \cdots, p.$
Atëherë vektori i katrorëve më të vegjël

$$\bar{b} = (b_1, \cdots, b_p)' = \boldsymbol{A}^{-1}\boldsymbol{g}$$

dhe për më tepër

$$Cov(\bar{b}) = \sigma^2\boldsymbol{A}^{-1}, \text{ dhe } cov(\bar{y}, b_i) = 0, \quad \text{për } i = 1, \cdots, p$$

2.3.2 Interpretimi gjeometrik i zgjidhjes me Metodën e Katrorëve më të Vegjël

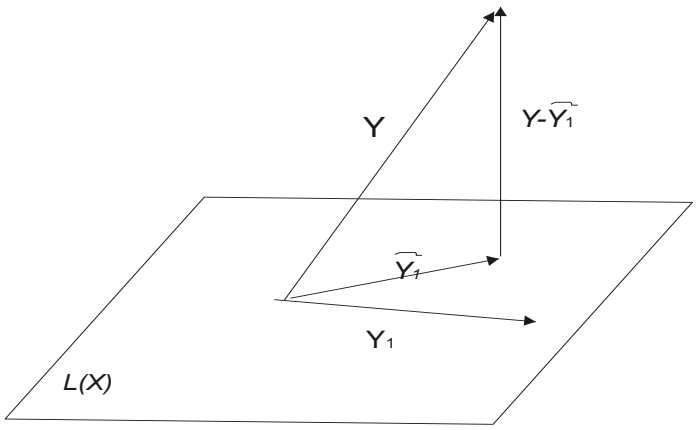
Në thelb të kësaj metode, qëndron përcaktimi i β , që të

$$\min_{\beta \in R^P} S(\beta) = \min_{\beta \in R^P}(\underline{Y} - X'\beta)'((\underline{Y} - X'\beta) = \min_{\beta \in R^P}(\underline{Y} - Y_1)'(\underline{Y} - Y_1)$$

ku $\underline{Y}$ janë shënuar vlerat e matura të Y
dhe me $Y_1 = X^T\beta$,

Ky minimum kërkohet për çdo $\beta \epsilon R^P$, ndërsa $Y_1=X^T \beta$ në fakt, nuk mbushin të gjithë R^n. Ato janë vetëm mbështjellset e shtyllave të X. Për këtë bashkësi mbështjelljesh, përdorim shënimin L(x).

Figura 2–1 Paraqitja gjeometrike



$(\underline{Y} - Y_1)$, gjeometrikisht na paraqet vektorin diferencë të dy vektoreve Y dhe Y_1.

Me këto simbolika të futura, përcaktimi i β kthehet në përcaktimin e vektorit diferencë $(\underline{Y} - Y_1)$ që ka gjatësi më të vogël, dhe Y_1 për këtë vektor është projeksioni i Y në hiperplanin L(x).

Siç e dimë ky vektor ka gjatësi më të vogël nëse ai është pingul me hiperplanin e Y_1, pra me L(x),që, siç e sqaruam më lart, është pjesë e R^n, ku ndodhen tërë $X^{'}\beta$.

Kështu në bazë të njohurive gjeometike, për vektorin ku arrihet minimumi i dëshiruar (po e shënojmë me $Y - \widetilde{Y}_1$), nga kushti i ortogonalitetit, mund të shkuajmë:

$$X(Y - \widetilde{Y}_1) = 0$$

$$XY - X\widetilde{Y}_1 = 0$$

$$XY = X\widetilde{Y}_1 = X(X^{'}\beta)$$

Vlerësimi *b* për parametrat β do të merret nga zgjidhja e ekuacionit matricor $XY = X(X^{'}\beta)$

Ndoshta ky vëzhgim gjeometrik i problemit mund të duket, i vështirë kur flitet për ndërtimin e një modeli të shumëfishtë. Por, në rastin kur $X \epsilon R^2$, pra në rastin e modelit linear të thjeshtë, ai ka paraqitjen gjeometrike si në figurën e mësipërme.

2.3.3 Vlerësimi intervalor i parametrave te regresit (Hapi II)

Programet e plota të posaçme kompjuterike, perveç vlerësimit pikësor te parametrave $\beta_0, \cdots, \beta_p$ me metoden e të katrorëve më të vegjël , japin informacion edhe për intervalet e besimit dhe për të kontrollin e hipotezava rreth këtyre parametrave.
Programet llogarisin gabimet standarde të koefiçientëve. Për çdo β_i, gabimi standard i koefiçientit, $se(b_i)$ është vlerësimi i devijimit standard të vlerësimit b_i të parametrit $\beta_i, i = 1, \cdots, p$. Meqënëse β është marrë si kombinim linear i y_i , që supozohet se kanë shpërndarje normale, atëherë mund të shfrytëzojmë faktin që $(\beta i$ –bi)/se_{bi} ka shpërndarje Studenti me n-p-1 shkallë lirie.
Kështu që intervali i besimit me koefiçient besimi $(1-\alpha) = \gamma$ të β_1 është:
$]L1, L2[_{\beta i} =]b_i - se(b_i)t_{1-\alpha}(v_R);\ bi + se(b_i)t_{1-\alpha}(v_R)[$ për $i = 1, \cdots, p$

2.3.4 Kontrolli i hipotezave (Hapi III)

Kontrolli i hipotezave për koefiçientët $\beta_1, \cdots, \beta_p$ ndahet në tri kategori: mund të testohen të gjithe koefiçientët $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, ose mund të testohet që secili koefiçient $\beta_k = 0, k = 1, \cdots, p$, ose mund të testojmë që çdo nënbashkësi me m koefiçientë është e barabartë me zero. Të gjitha këto hipoteza mund të interpretohen nëpërmjet variablave të pavarur si më poshtë.
Së pari kontrollojmë hipotezën $H_{0:}\ \beta_1 = \beta_2 = \ldots = \beta_p = 0$
(pra praktikisht që variablat x_1, $x_{2,\ldots}$ x_p s'ndikojnë në vlerësimin e y)
Kjo hipotezë s'është gjë tjetër veçse hipoteza "variablat e pavarur $X_1, \cdots, X_p$ nuk e përmirësojnë parashikimin e Y më shumë se $\hat{y} = \overline{y}$",
me hipotezë alternative " jo të gjithë koefiçientët janë të barabartë me zero"
Nëse nuk hidhet poshtë kjo hipoteze do të thotë që, pranohet $\overline{y}$ si parashikuesi më i mirë i Y, dhe nuk ia vlen ndërtimi i modelit të regresit linear.

Për kontrollin e kësaj hipoteze përdorim statistikën

$$\frac{MS_D}{MS_R} = \frac{\sum (\bar{y} - \tilde{y})^2 \Big/ p}{\sum (y_i - \tilde{y})^2 \Big/ (n - p - 1)}$$ me shpërndarje F(p,n-p-1)

Për të spjeguar këtë, mjafton të kujtojmë se $SS_R/\ \sigma^2$ ka shperndarje χ^2 (n-p-1),
dhe $SS_T/\ \sigma^2 = \sum (y_i - \bar{y})^2 / \sigma^2$ ka shpërndarje χ^2 (n- 1).

Si përfundim $SS_{D/}\ \sigma^2 = \sum (\tilde{y} - \bar{y})^2 / \delta^2$ do të ketë shpërndarje χ^2 (p).

Nëse hipoteza H_0 pranohet si e vërtetë, kjo do të thotë se për vlerësimin e Y mund të përdorim $\bar{y}$. Nëse kjo hipotezë nuk pranohet si e vërtetë, si e tillë pranohet hipoteza alternative, Nëse hipoteza H_0 hidhet poshtë atëhere do të pranohet hipoteza alternative si e vërtetë, pra që jo të gjithë parametrat janë 0 ose njëlloj, "disa nga variablat e pavarur përmirësojnë parashikimin e Y më mirë se $\hat{y} = \overline{y}$".

Ndaj kalohet në etapën tjetër.
Etapa e dytë: Shtrohen hipotezën H_0: $\beta_k=0$, për k=1..p, me hipotezë alternative H_1: $\beta_k \neq 0$, Hipotezat $H_0: \beta_k = 0$ për $k = 1, \cdots, p$ mund të shihet si hipoteza ku "varabli X_k nuk përmirëson parashikimin e Y krahasuar me atë që fitohet nga efekti në Y i $p-1$ variablave të tjerë" Gjeometrikisht mendohet që hiperplani që po ndërtohet është paralel me këtë bosht.
Të dhënat për kontrollin e hipotezës së shtruar përcaktohen nga raporti F që jepet në kolonën e fundit të tabelës së analizes së variancës, dhe është,

$$F = \frac{MS_D}{MS_R}.$$

Kur është e vërtetë H_0, F ka njëshpërndarje F me $v_D = p$ dhe $v_R = n - p - 1$ shkallë lirie. p-vlera është sipërfaqja në të djathtë të F të vrojtuar, nën grafikun e shpërndarjes së $F(v_D, v_R)$.
Si test statistikor për kontrollin e hipotezes merret:

$$F = \frac{{b_k}^2}{[se(b_k)]^2}$$

e cila kur H_0 është e vërtetë, ka një shpërndarje F me 1 dhe $v_R = n - p - 1$ shkallë lirie. p-vlera është hapësira në të djathtë të F në grafikun e shpërndarjes $F(1, v_R)$.
Disa programe kompjuterike e japin këtë vlerë të F për çdo b_k.
Programe të tjerë(SPSS) japin vlerën e statistikës ekuivalente

$$t = \frac{b_k}{se(b_k)}$$

e cila, nëse H_0 është e vërtetë ka shpërndarjen Studenti me $v_R = n - p - 1$ shkallë lirie. p-vlera është dyfishi i hapësirës në të djathtë të $|t|$ në grafikun e S(v_R).

Etapa e tretë: Hipoteza e fundit, që një nënbashkësi prej m koefiçientësh është zero është më e vështirë të provohet. Pa bërë ndonjë kufizim të madh, le të mendojmë që kjo nënbashkësi konsiton ne m koefiçientët e parë $\beta_1, \cdots, \beta_m$. Atëherë të provosh $H_0: \beta_1 = \cdots = \beta_m = 0$ është e njëjtë sikur të provosh që "m variablat $X_1, \cdots, X_m$ nuk permirësojnë parashikimin e Y nga ajo që fitohet nga regresi i Y nga $X_{m+1}, \cdots, X_p$".
Fillimisht, ekzekurohet programi për të llogaritur regresin e Y me $X_{m+1}, \cdots, X_p$. Nga analiza e tabelës së variancës marrim shumën e katrorëve të mbetjes prej ${SS_R}'$. Më pas përdoret programi për të llogaritur regresin e Y me $X_1, \cdots, X_m, \cdots, X_p$, dhe nxjerrim shumën e katrorëve te mbetjes dhe katrorin mesatar prej respektivisht SS_R dhe MS_R. Atëherë testi statistikor për H_0 është

$$F = \frac{({SS_R}' - SS_R)/m}{MS_R}$$

e cila kur H_0 është e vërtetë ka shpërndarje F me m dhe $v_R = n - p - 1$ shkallë lirie. p-vlera është hapësira në të djathtë të F së vrojtuar nën grafikun e shpërndarjes $F(m, v_R)$.

2.3.5 Vërejtje

1. Varianca e $\hat{y}$ prej $x_1, \cdots, x_p$ është

$$V(\hat{y}) = \sigma^2 \left(\frac{1}{n} + d' A^{-1} d\right)$$

ku elementët e A janë përcaktuar në seksionin2.3.1 nga(*), dhe
$d = (x_1 - \bar{x}_1, \cdots, x_p - \bar{x}_p)'$, me $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ik}$.
Një interval besimi me koefiçient besimi $(1-\alpha)$, për mesataren e vlerës së vërtetë të Y në se $x_1, \cdots, x_p$ njihet, është

$$\hat{y} \pm \left[s^2\left(\frac{1}{n} + d'A^{-1}d\right)\right]^{\frac{1}{2}} t_{1-\left(\frac{\alpha}{2}\right)}(n-p-1)$$

Një interval besimi me koefiçient besimi $(1-\alpha)$ për një vlerë të vetme të parashikuar të Y nëse $x_1, \cdots, x_p$njihet, është

$$\hat{y} \pm \left[s^2\left(1 + \frac{1}{n} + d'A^{-1}d\right)\right]^{\frac{1}{2}} t_{1-\left(\frac{\alpha}{2}\right)}(n-p-1)$$

Disa programe shfaqin elementët e A^{-1} për të lehtësuar llogaritjen e ekuacioneve të mësipërme.
2. Nëse testojmë Hipotezat $H_0: \beta_k = 0$ për disa vlera të k me të njëjtin nivel rëndësie α, atëherë niveli i rëndësisë së përbashkët nuk është patjetër α. Për të shmangur këtë problem mund të ndërtojmë intervalet e besimit së shumëfishtë për të gjitha $\beta_k, k = 1, \cdots, p$, në mënyrë që koefiçienti i besimit të përbashkët të jetë $1-\alpha$. Intervali i besimit të shumëfishtë me koefiçient $(1-\alpha)$ për $\beta_k$është

$$b_k \pm (se(b_k)[pF_{1-\alpha}(p, n-p-1)]^{1/2}$$

Hipoteza $H_0: \beta_k = \beta_k^{(0)}$ bie poshtë me nivel rëndësie α, nëse $\beta_k^{(0)}$ nuk është në këtë interval.

2.3.6 Përfundime

- Në regresin e shumëfishtë standart të gjithë variablat e pavarur futen njëkohësisht në model.
- Vlerësimi i R and R² përcaktojnë fortësinë e lidhjes mes variablave të pavarur dhe variablit të varur. Testi Fisher përdoret për të përcaktuar nëse kjo lidhje e përcaktuar nga zgjedhja mund të përgjithsohet, apo jo, për të gjithë popullimin .
- Testi t përdoret për të vlerësuar lidhjen individuale mes çdo variabli të pavarura dhe variablit të varur.

2.4 Zbatim 1

Nëse i referohemi zbatimit **"Efekti i SO_2 në prodhimet bujqësore**", vihet re perdorimi i kësaj metode studimi të të dhenave. Në ketë zbatim kërkohet te kontrollohet nëse ndotja industriale ndikon apo jo ne prodhueshmërine e produkteve bujqesore. Pra a ndikon rritja e produkteve në zonen industriale apo jo industrale, në sasine e produktit të marrë?
Për këtë qëllim janë kontrolluar në të dy zonat, sasia e prodhimit të të njëjtave produkte bujqesore. Pas përdorimit të testeve të përshtatshme, është vënë re se për produktet 'grurë', *'misër'* dhe *'fruta*', nuk ka dallim midis zonave. Por, sasia e produkteve bujqësore dihet që varet nga përkujdesja ndaj sipërfaqeve të

shfrytëzueshme per të mbjella dhe kushtet abiotike e biotike. Për të studiuar efektin e këtyre faktorëve në këtë problem para aplikimit të metodës ne fjalë është kontolluar :
A është bashkësia e të dhënave e përshtatshme për këtë aplikim statistikor?
Kontrollohet më parë që te mos ketë probleme me mungesën e të dhënave, probleme me outliers, plotësohen kushtet e aplikimit të kesaj metode dhe më pas, që ndarja e të dhenave që shfrytëzohen, të konfirmojë per besueshmeri te rezultateve.
Variablat 'distance', që përcakton distancën në metër te individit nga pusi, *'mbetje urbane'* që kodohet sipas shkallës së trajtimit të mbetjëve urbane, në rastin tonë më i lartë kodi më i ulët niveli i trajtimit të kesaj lloj mbetje; dhe kujdesi ndaj '*ujrave të zeza'*, që eshtë koduar në mënyrë të tillë që rritja e kodit tregon përkujdesje të ulet të trajtimit të tyre, janë konsideruar pas studimit, si variablat që duhen marrë në shqyrtim.

Metoda e regresit të shumëfishtë kerkon që variabli i varur *'bimet*' të jetë metrik, dhe variablat e pavarur, të jenë metrikë ose kategorikë, që përcaktojnë parametrat ne fjalë. *'ujrat e zeza'* dhe *'sistemi i mbeturinave'* jane variabla ordinal të përcaktuara sipas niveleve, pra në një farë mënyre mund të mendohet qe keto variabla mund të masin parametrat në fjale. Mendohet që variablat e ketij lloji mund të trajtohen si variabla metrikë. Por, ka statisticienë që nuk janë dakord me këtë përdorim. Ndaj, në fund të interpretimit duhet bërë një vërejtje, nëse këto lloj variablash përdoren.
'distanca nga pusi 'është variabël metrik ndaj ai përdoret në metodën e regresit linear.
Minimumi i raportit të vëllimit të zgjedhjes që shfrytëzohet me variablat e pavarur që marrin pjesë në studim, duhet të jetë minimumi 5 me 1. Por 155 rastet e vrojtuara dhe 3 variabla të pavarur bëjnë që raporti i kësaj analize të jetë 51.7 me 1, duke plotesuar kështu edhe kushtin e preferuar15 me 1.

Tabela 2–1 Tabela e koefiçientëve të modelit të regresit

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	731.418	293.331		2.493	.014
	area	-76.385	116.533	-.157	-.655	.513
	sistemi i ujrave te zeza	-215.809	56.983	-.298	-3.787	.000
	terheqja e mbeturinave	-7.930	21.784	-.034	-.364	.716
	distanca shtepi-pusne meter	-.053	.183	-.065	-.288	.774

a. Dependent Variable: fruta

Tabela 2–2 Afishimet e t-testeve dhe koeficienteve te regresit

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	929.111	287.930		3.227	.002
	area	-20.027	114.388	-.043	-.175	.861
	sistemi i ujrave te zeza	-150.558	55.933	-.216	-2.692	.008
	terheqja e mbeturinave	18.431	21.383	.083	.862	.390
	distanca shtepi-pusne meter	-.022	.180	-.027	-.120	.905

a. Dependent Variable: miser

Ndikimi i faktorit abiotik *ujrat e zeza* duhet marrë parasysh, sepse tek *gruri* (t-test -3.415 dhe sig. 0.01), *misri* (t-test -2.692 dhe sig. 0.08) dhe *frutat* (t-test -3.787 dhe sig.< 0.01). Vlerat negative të koefiçientëve përkatësisht -436, -150.5 dhe -215 në modelin e regresit linear përcaktojnë drejtimin e lidhjes. Duke verejtur kodin e variablit *ujra të zeza*, ndërsa kodi rritet nuk ka asnje proces sanitar në trajtimin e ujrave të zeza, arrihet në përfundimin se një sistem më i mirë i ujrave të zeza ndimon në prodhueshmëri me të lartë tek disa bime.
Hipotezat zero që nuk kanë lidhje gjithë variablat e pavarur me variablin e varur, *prodhueshmëria e bimës*, bie poshtë . Ne tabele (model summary) janë paraqitur dhe vlerat e statistikes R^2- change. Informacioni i marrë nga variablat e shtuar, redukton parashikimin në prodhueshmerinë e bimëve (7.6% për *grur*in, 5.4% per *misr*in, and 6.7% për *frutat*).

2.4.1 Vërejtje

Në modelin e regresit, β_i përcakton shkallën e ndryshimit të Y në varësi të X_i, kur vlerat e $X_j, j = 1, \cdots p,\ j \neq i$ janë mbajtur të pandryshueshme. Megjithatë, këto koefiçientë mund të mos ketë kuptim t'i krahasojmë në madhësi për shkak të diferencave në njësitë e matjes së $X_1, \cdots, X_p$. Kjo vështirësi mund të kalohet duke përdorur *variabla të pavarur të standartizuar*[5]. Kështu, për $j = 1, \cdots, p$, le të jenë $Z_j = X_j/s_j$, ku ${s_j}^2 = \sum_{i=1}^{n}\left(x_{ji} - \bar{x}_j\right)^2/(n-1)$.
Modeli i regresit të shumëfishtë në termat e Z_j jepet tashmë prej ekuacionit

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \cdots + \gamma_p z_{pi} + e_i, \qquad i = 1, \cdots, n$$

ku $\gamma_k, k = 0 \cdots, p$, janë parametra të panjohur dhe e_i janë terma gabimi të pavarura $N(0, \sigma^2)$. Vlerësuesit e katrorëve më të vegjël c_k tëγ_k dhe testet e hipotezave rreth γ_k ndjekin teorinë e mësipërme me z_j dhe γ_k që zëvëndësojnë x_j dhe β_k respektivisht. Avantazhi i këtij standartizimi është që $\gamma_1, \cdots, \gamma_p$ janë tani madhësitë që përcaktojnë ndryshimin, të matura jo më në shkallë të ndryshme. Kjo na lejon të arrijmë në konkluzionin rreth shkallës së kontributit të variablave të pavarur $Z_1, \cdots, Z_p$ (ose në

mënyrë të njëjtë, $X_1, \cdots, X_p$). Pra, një vlerë e lartë e c_j sjell një shkallë të lartë të kontributit të Z_j (ose X_j), $j = 1, \cdots, p$.

2.5 Koefiçienti i korrelacionit të shumëfishtë

Do diskutojmë teorikisht modelin e regresit linear të shumëfishtë. Teoria pranon që $p + 1$ variablat $Y, X_1, \cdots, X_p$ janë të gjithë variabla të rastësishëm, dhe për më tepër ata kanë një shpërndarje normale të shumëpërmasore. Do të tregohet këtu që mesatarja e shpërndarjes së kushtëzuar te Y, në se $X_1 = x_1, \cdots, X_p = x_p$ njihet, është funksion i regresit të shumëfishtë linear të $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. Kjo motivon modelin e regresit të shumëfishtë linear, në të cilin varianca e gabimit σ^2 është funksion i variancës ${\sigma_y}^2$ të Y dhe i një madhësie që quhet koefiçient i korrelacionit të shumëfishtë.

Shqyrtojmë shpërndarjen normale të shumëfishtë të $Y, X_1, \cdots, X_p$ me mesatare $\mu_y, \mu_1, \cdots, \mu_p$ dhe varianca ${\sigma_y}^2, {\sigma_1}^2, \cdots, {\sigma_p}^2$, respektivisht. Shënojmë kovariancën e Y me X_i me σ_{yi}, dhe kovariancën e X_i me X_j me σ_{ij} për $i, j = 1, \cdots, p$. Gjithashtu përcaktojmë koefiçientët e korelacionit të popullimit:

$\rho_{yx_i} = \frac{\sigma_{yi}}{\sigma_y \sigma_i}$ dhe $\rho_{x_i x_j} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$

Për vlera të dhëna të $X_1 = x_1, \cdots, X_p = x_p$, ekziston një nënpopullim i vlerave korresponduese të të Y. Shpërndarja e saj, shpërndarja e kushtëzuar e Y në se $X_1 = x_1, \cdots, X_p = x_p$, është gjithashtu normale me mesatare

$$\mu_{y,x_1 \cdots x_p} = \mu_y + \beta_1 (x_1 - \mu_1) + \cdots + \beta_p (x_p - \mu_p)$$

Kjo quhet dhe pritja matematike me kusht e Y, kur dihet se $X_1 = x_1, \cdots, X_p = x_p$, ose e regresit të Y ndaj $X_1, \cdots, X_p$. Vlerat $\beta_1, \cdots, \beta_p$ quhen koefiçientë të regresit të pjesshëm, dhe janë funksion i variancave dhe kovariancave të mësipërme. Varianca e kësaj shpërndarje të kushtëzuar është

${\sigma}^2 = {\sigma_y}^2 (1 - \rho^2_{y,x_1 \cdots x_p})$ (*)

ku $\rho_{y,x_1 \cdots x_p}$, rrënja katrore positive e $\rho^2_{y,x_1 \cdots, x_p}$, quhet koefiçienti i korrelacionitt të shumëfishtë te Y ndaj $X_1, \cdots, X_p$.

Nëse përcaktojmë variablin e rastësishëm $e = Y - \mu_{y,x_1 \cdots x_p}$, atëherë shpërndarja me kusht e e -së, nëse dihet se $X_1 = x_1, \cdots, X_p = x_p$, është $N(0, \sigma^2)$. Bazuar në këtë, mund të shkruajmë

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e$$

ku

$$\beta_0 = \mu_y - \beta_1 \mu_1 - \cdots - \beta_p \mu_p$$

dhe e ka shpërndarje $N(0, \sigma^2)$. Vihet re se ekuacioni i mëparshëm ka të njëjtën formë me modelin e regresit të shumëfishtë linear të trajtuar më parë.

2.5.1 Përfundime

1. Koefiçienti i korelacionit të shumëfishtë $\rho_{y,x_1,\cdots,x_p}$përcakton masën e lidhjes lineare ndërmjet Y dhe bashkësisë së variablave $\{X_1,\cdots,X_p\}$. Koefiçienti i korelacionit të shumëfishtë merr vlera $0 \leq \rho_{y,x_1,\cdots,x_p} \leq 1$. Një vlerë zero tregon që Y është i pavarur nga bashkësia $\{X_1,\cdots,X_p\}$, ndërsa një vlerë 1 tregon varësi perfekte lineare, pra që, Y mund të shprehet saktësisht si një kombinim linear i $X_1,\cdots,X_p$.
Për të analizuar fortësine e modelit të regresit linear të shumëfishtë, përdoret zakonisht ky rregull. Nëse koefiçienti i korelacionit është më e vogël se 0.2, lidhja e variablave sipas modelit te regresit konsiderohet shumë e dobët; nëse koefiçienti i korelacionit është më i madh se 0.2 dhe më i vogël se 0.4, atëhere thuhet se lidhja sipas modelit të regresit është e dobët, për vlerën e koefiçientit nga 0.4 në 0.6 modelimi quhet i moderuar. Modeli i regresit linear quhet i fortë dhe shumë i fortë përkatësisht në rastet kur ky koefiçient i korelacionit, i llogaritur si mësipërm, ka vlera përkatësisht nga 0.6-0.8 dhe nga 0.8 në 1.

2. Duke zgjidhur ekuacionin për koefiçientin e korelacionit të shumëfishtë, marrim:

$$\rho^2_{y,x_1\cdots x_p} = \frac{\sigma_y{}^2 - \sigma^2}{\sigma_y{}^2} \qquad (*)$$

Katrori i koefiçientit të korelacionit të shumëfishtë përcakton pjesën e variancës së Y që "spjegohet" nga relacionin e regresit me $X_1,\cdots,X_p$.
3. Koefiçienti i korelacionitt të shumëfishtë është jonegativ, me përkufizim. Kështu në rast të shpërndarjes normale dypërmasore, ku $(p = 1)$, kemi

$$\rho_{y,x_1} = \rho_{x_1,y} = |\rho_{x_1 y}|$$

ku $\rho_{x_1 y}$është koefiçienti i korelacionit të thjeshtë ndërmjet X_1 dhe Y.
4. Kur $p = 2$, arsyetimet e mësipërme mund të shkruhen:

$$\mu_{y,x_1x_2} = \mu_y + \beta_1(x_1 - \mu_1) + \beta_2(x_2 - \mu_2)$$

Grafiku i këtij ekuacioni është një plan (që quhet plani i regresit të Y sipas X_1 dhe X_2) në hapësirën e përcaktuar nga boshtet koordinative x_1 , x_2, dhe nga μ_{y,x_1x_2}. Grafku i ekuacionit për $p > 2$ paraqet një hiperplan në hapësirën $p + 1$ dimensionale të përcaktuar nga $x_1,\cdots,x_p$ dhe nga vlera $\mu_{y,x_1\cdots x_p}$.
5. Korelacioni I shumëfishtë është maksimumi i një korelacioni të thjeshtë mes Y dhe kombinimit linear të $X_1,\cdots,X_p$. Për më tepër, $\mu_{y,x_1\cdots x_p}$është kombinimi linear që siguron këtë korelacion maksimal. Lidhja ndërmjet koefiçientit të korelacionit të shumëfishtë dhe parametrave të regresit $\beta_1,\cdots,\beta_p$ do të diskutohet më vonë.
6. Koefiçienti i korelacionit të shumëfishtë nuk varet nga shkalla e matjes se ndryshoreve të rastit të $Y, X_1,\cdots, ose\ X_p$.
7. Koefiçienti i korelacionit të shumëfishtë mund të shfrytëzohet edhe në një formë tjetër. Kështu, $(1 - \rho^2_{y,x_1\cdots x_p})^{1/2}$ është pjesa e i devijimit standard të Y që mbetet "e pashpjeguar" nga ndryshimet e $X_1, X_2,\cdots, X_p$. Për shembull, një koefiçient i korrelimit të shumëfishtë 0.9, në dukje shumë i lartë, lë akoma 44% të devijimit standard të Y të pa shpjeguar.

2.5.2 Zbatim 1 (vazhdim)

Nëse i referohemi përsëri **"Efekti I SO_2 në prodhimet bujqësore**", pasi u vu re se për produktet '*grurë*', *'misër'* dhe *'fruta*', nuk kishte dallim midis zonave, u kontrollua nëse sasia e produkteve bujqësore varet nga kushtet abiotike e biotike. Variablat '*distance*', që përcakton distancën në metër te individit nga pusi, *'mbetje urbane'* që kodohet sipas shkallës së trajtimit të mbetjëve urbane, në rastin tonë më i lartë kodi më i ulët niveli i trajtimit të kesaj lloj mbetje; dhe kujdesi ndaj '*ujrave të zeza'*, që eshtë koduar në mënyrë të tillë që rritja e kodit tregon përkujdesje të ulet të trajtimit të tyre, u trajtuan me metoden e regresit linear të shumëfishte dhe u mor modeli I regresi te përcaktuar sipas tabelës 2-2.

Tabela 2–3 Tabela e vlefshmerisë së modelit për 'frutat'

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.301[a]	.091	.067	213.48478	.091	3.743	4	150	.006

a. Predictors: (Constant), distanca shtepi-pusne meter, sistemi i ujrave te zeza, terheqja e mbeturinave, area

Tabela 2–4 Tabela e vlefshmerisë së modelit për 'misri '

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.232[a]	.054	.028	209.55396	.054	2.126	4	150	.050

a. Predictors: (Constant), distanca shtepi-pusne meter, sistemi i ujrave te zeza, terheqja e mbeturinave, area

Afishimet, tabela 2.3 dhe 2.4, tregojnë se kemi të bëjme me modele regresi që duhen marrë parasysh, (p-vlerat për të dy modelet janë 0.006 dhe 0.05) . Por nëse përqëndrohemi tek vlerat e afishuara, vleresimet për R janë 0.35 dhe 0.24, të cilat në baze të vërejtjes 1, flasin per nje model të dobët regresi. Pikerisht modelet e shqyrtuara justifikojnë 11% dhe 5.5% të ndryshimeve në prodhushmërinë e bimeve në fjalë.

2.6 Koefiçienti i korelacionit të pjesshëm

Këtu diskutohet një tjetër koefiçient korelacioni, koefiçienti i korelacionit të pjesshëm. Ky përdoret për të matur lidhjen lineare ndërmjet çfarëdo dy prej $p+1$ variablave $Y, X_1, \cdots, X_p$, pas heqjes së 'efektit' të një nënbashkësie, jo boshe, të $p-1$ variablave të mbetur. Në veçanti, ai përcakton varësinë ndërmjet Y dhe një variabli të pavarur X_m, pas heqjes nga Y, të varësisë lineare të një nënbashkësi të k prej $p-1$ variablave të pavarur të mbetur $X_i, i = 1, \cdots, p,\ \ i \neq m$. Kjo varësi lineare e Y ndaj nënbashkësinë së k variablave quhet 'efekt' i nënbashkësisë në fjalë, tek ndryshorja e rastit Y. Teoria e koefiçientit të korelacionimit të pjesshëm lidhet me shpërndarjen me kusht.

2.6.1 Shtrimi i problemit

Le të jenë l dhe h, dy prej variablave çfarëdo nga $Y, X_1, \cdots, X_p$, dhe le të jetë c një nënbashkësi jo boshe e $p-1$ variablave të mbetur. Përcaktohen ndryshoret e rastit $Z_1 = l - \mu_{l,c}$ ku $\mu_{l,c}$ pritja me kusht e l, për c e dhënë dhe, $Z_2 = l - \mu_{h,c}$, ku $\mu_{h,c}$ pritja me kusht e h, për c e dhënë. Duhet patur parasysh se Z_1 dhe Z_2 janë ndryshore rasti, sepse ato janë funksione të ndryshoreve të rastit në c. Atëherë koefiçienti i korelacionit të pjesshëm të l dhe h, për c e dhënë është:

$$\rho_{lh,c} = \rho_{z_1 z_2}$$

ku $\rho_{z_1 z_2}$ është koefiçienti i zakonshëm i korelacionit ndërmjet Z_1 dhe Z_2.

Shfaqen zakonisht dy raste të veçanta.

Një rast është kur $l = Y, h = X_m, m = 1, \cdots, p,$ dhe c është nënbashkësia e $p-1$ variablave të pavarur të mbetur. Ky korelacion i pjesshëm do shënohet me $\rho_{y,x_m,c}$.

Rasti i dytë është kur $l = Y, h = X_m$, dhe c është nënbashkësi e k variablave të parë të pavarur $\{X_1, X_2, \cdots, X_k\}$, ku $l \leq k < m \leq p$. Koefiçienti i korelacionit të pjesshem në këtë rast, do të shënohet me $p_{yx_m, x_{1\cdots x_k}}$. Në përgjithësi në se janë k variabla në c, atëherë koefiçienti i korelacionit të pjesshëm thuhet se është i $rendit\ k$.

2.6.1.1 Përfundime

1. Koefiçienti i korelacionimit të pjesshëm $\rho_{lh,c}$ është percaktues I lidhjes lineare ndërmjet l dhe h, kur vlerat e variablave të paraqitura c s' ndryshojnë. Ajo merr vlera nga -1 në 1; një vlerë zero tregon që l është e pavarur nga h, kur c mbahet e pandryshuar.
2. Përdoret shumë një identitet ndërmjet koefiçientëve të korelacionimit të pjeshëm dhe të shumëfishtë për bashkësinë e variablave $Y, X_1, \cdots, X_{k-1}, X_k,\ \ k = 2, \cdots, p$, që është

$$1 - \rho^2_{y,x_1 \cdots x_k} = (1 - \rho^2_{y,x_1 \cdots x_{k-1}})(1 - \rho^2_{yx_k, x_1 \cdots x_{k-1}})$$

Që do të thotë se

$$V(Y|X_1, \cdots, X_k) = V(Y|X_1, \cdots, X_{k-1})\left(1 - \rho^2_{yx_k, x_1 \cdots x_{k-1}}\right)$$

ku $V(Y|X_1,\cdots,X_i)$është variance (dispersioni)me kusht e Y, për $X_1,\cdots,X_i$, $i=1,\cdots,p$ të dhënë. Kështu

$$\rho^2_{yx_k,x_1\cdots x_{k-1}}=\frac{V(Y|X_1,\cdots,X_{k-1})-V(Y|X_1,\cdots,X_k)}{V(Y|X_1,\cdots,X_{k-1})}$$

Kështu që katrori i koefiçientit të korelacionit të pjeshëm është në propocional me variancën e Y "që shpjegohet" nga shtesa e X_k tek bashkësia $\{X_1,\cdots,X_{k-1}\}$.

3. Kemi gjithashtu

$$\rho_{yx_m,c}=\beta_m\left[\frac{V(X_m|c)}{V(Y|c)}\right]^{1/2},\qquad m=1,\cdots,p$$

ku c është tërë nënbashkësia e $p-1$ variablave të mbetur dhe $V(X_m|c)$është varianca me kusht e X_m për varibla e dhëna c. Kështu të provosh që $\beta_m=0$ është e njëjtë sikur të provosh që $\rho_{yx_m,c}=0$.

4. Korelacioni i pjesshëm mund të llogaritet nga një proces rekrusiv si më poshtë. Në se l,h dhe d paraqesin çdo tre varibla të marrë në bashkësinë$\{Y,X_1,\cdots,X_p\}$, atëherë të gjitha korelacionet e pjesshme të rendit të parë jepen prej

$$\rho_{lh,d}=\frac{\rho_{lh}-\rho_{ld}\rho_{hd}}{\sqrt{(1-\rho^2_{ld})}\sqrt{(1-\rho^2_{hd})}}$$

ku afishimi në të djathtë paraqet korelacionet e thjeshta. Hap pas hapi, duke përdorur formulën rekursive

$$\rho_{lh,cd}=\frac{\rho_{lh,c}-\rho_{ld,c}\rho_{hd,c}}{\sqrt{\left(1-\rho^2_{ld,c}\right)}\sqrt{\left(1-\rho^2_{hd,c}\right)}}$$

ku c është çdo nënbashkësi e variblave të mbetur, mund të fitojmë koefiçientët e korelacionit të pjesshëm të çdo rendi.

2.6.2 Vlerësimi dhe testimi i hipotezave për koefiçientët e korelacionit të shumëfishtë dhe të pjesshëm

Shqyrtohet tani problemi I llogaritjes dhe interpretimit i koefiçientëve të korelacionit nisur nga shfrytezimi i bazës së të dhënave. Me qenë se teorikisht kërkohet që të gjitha $p+1$ variblat të jenë variabla të rastit, supozohet se vlerat e vrojtuara $\left(y_1,x_{11},\cdots,x_{p1}\right),\cdots,\left(y_n,x_{1,n},\cdots,x_{pn}\right)$ janë fituar nga zgjedhja e rastësishme e n individëve nga popullimi normal shumëpërmasor. Për çdo individ, të gjithë $p+1$ variablat janë matur njëkohësisht. Vlerësuesit e mesatareve, variancave dhe kovariancave të këtij popullimi merren nga vlerësimi i mesatareve, variancave dhe kovariancave të zgjedhjes Lehtësisht softet statistikore bejne llogaritjen e mesatares, variancës dhe kovariances së vektorit të të dhënave, permes analizës deskriptive. Madje ato llogarisin edhe matricën e kovariancës midis $p+1$ variblave.

U vu re se që shpërndarja me kusht e Y, për $X_1=x_1,\cdots,X_p=x_p$ të dhënë, çon në modelin e regresit linear të shumëfishtë. Kështu rastin tonë kemi:

$$y_i=\beta_0+\beta_1x_{1i}+\cdots+\beta_px_{pi}+e_i,\qquad i=1,\cdots,n$$

me e_i të pavarur $N(0,\sigma^2)$. Me qenë se ky ekuacion është identik me ekuacionin e regresit të diskutuar në seksionin e mëparshëm, atehere vlerësuesit për $\beta_0,\beta_1,\cdots,\beta_p$,

dhe σ^2 dhe hipotezat e testit të diskutuara në atë seksion janë të vlefshme këtu. Ato që mbeten për t'u vlerësuar janë koefiçientët e korrelacionit të pjesshëm dhe të shumëfishtë.

Diskutojmë tani vlerësimin e koefiçientëve të korelacionit të shumëfishtë dhe atij të pjesshëm. Vlerësuesi i koefiçientit të korrelimit të shumëfishtë, i shënuar si $r_{y,x_1,\cdots,p}$, zakonisht afishohet nga programet statistikore me shënimin $R\ i\ shumefishte$ ose Koefiçient i korelacionit të shumëfishtë. Ai mund të llogaritet gjithashtu në analizën e tabelës së variancës , nga :

$$r_{y,x_1\cdots x_p} = +\sqrt{\frac{SS_D}{SS_T}}$$

Ai mund të përcaktohet edhe si rrënja katrore, pozitive, e koefiçientit të përcaktimit R^2. Ky vlerësim pikësor për koefiçientin e korelacionit të shumefishtë është gjithnjë jo negativ.

Për cfarë vlen vlerësimi për koefiçientin e korelacionit?

Së pari, ai përcakton masën e varësisë lineare të Y nga të gjithë variablat e pavarur. Për të kontrolluar hipotezën bazë se, nuk ka varësi lineare , pra që, $H_0: \rho_{y,x_1\cdots x_p} = 0$, mund të përdorim të rezultatet e $F - testit$, me qenë se hipoteza e shtruar është ekuivalente me $H_0: \beta_1 = \cdots = \beta_p = 0$. Mund të përdoret gjithashtu edhe $F - testi$ ekuivalent:

$$F = \frac{n-p-1}{p}\frac{r^2_{y,x_1\cdots x_p}}{1-r^2_{y,x_1\cdots x_p}}$$

p-vlera është sipërfaqja në të djathtë të F nëndensitetin probabilitar të shpërndarjes $F(p, n-p-1)$.

Së dyti, thamë se katrori i këtij koefiçienti të vlerësuar përcakton "pjesën e variacionit të Y që shpjegohet nga regresit linear i Y ndaj $X_1, \cdots, X_p$".

Vëzhgimet që mungojnë, në rastin shumëpërmasor

Në analizat që janë me një variabël (p.sh. testet t), trajtimi më arsyeshëm i rastit kur vlera mungon te X (variabli që do analizohet) është heqja e këtij vrojtimi nga zgjedhja. Situata është e ndryshme kur kemi të bëjme me analiza që janë të natyrës shumëpërmasore, p.sh. kur p variablat $X_1, \cdots, X_p$ vëzhgohen në çdo vrojtim. Në këtë situatë në se një prej variablave u mungon një vlerë përshembull X_1, nuk është e nevojshme të heqim këtë vrojtim nga analiza dhe të humbnim informacionin për $X_2, \cdots, X_p$. Me që ekuacioni i analizës së regresit të shumëfishtë sikurse edhe procedurat multivariate janë funksione të vektorit mesatar μ dhe të matricës së kovariancës $\sum$, mund ta lëmë këtë rast në zgjedhje dhe të përdorim informacion që kemi në llogaritjen e vlerësimit të vektorit mesatar $\overline{x}$ dhe matricës së kovariancës S.

Tani të diskutojmë metoda të ndryshme për vlerësimin e μ dhe $\sum$ (apo njëlloj të matricës së korrelimit R), kur mungojnë disa vlera. Le të jenë n_i numri i rasteve kur $X_i$është pa mungesa, n_{ij} numri i rasteve kur të dyja X_i dhe X_j janë pa mungesa, dhe n_c numri i rasteve kur të gjitha $X_1, \cdots, X_p$janë pa mungesa ($n_i, n_{ij}, n_c \leq n$ - vëllimi I zgjedhjes, $i, j = 1, \cdots p, i \neq j$). Atëherë disa metoda për të përcaktuar $\overline{x}$ dhe S (ose $\hat{R}$) janë si më poshtë:

Metoda 1
Vlerësimi i $\overline{x}$ dhe S duke përdorur të gjitha rastet e n_c. Kjo quhet metoda e eleminimit të rastit.

Metoda 2
Përdorimi i n_i vëzhgime e llogaritet $\bar{x}_i$. Zëvendësojmë $\bar{x}_i$ për çdo observim të munguar në X_i. Më pas përdoret vrojtimi i plotësuar i n vëzhgimeve, për të marrë $\overline{x}$ dhe S. Kjo quhet metoda e zëvendësimit mesatar.

Metoda 3
Përdoren n_i vëzhgimet e llogaritet $\bar{x}_i$ dhe ${s_i}^2$, dhe përdoren i n_{ij} observimeve për të fituar s_{ij}. Këto statistika përcaktojnë $\overline{x}$ dhe S.

Metoda 4
Përdoren n_i vëzhgimet e llogaritet $\bar{x}_i$ dhe ${s_i}^2$, dhe përdoren i n_{ij} vëzhgimet e llogaritet r_{ij}. Më pas s_{ij} llogaritet si $s_{ij} = r_{ij}(s_i)(s_j)$, që është i vetmi ndryshim nga metoda 3. Metodat 3 dhe 4 quhen dhe metodat e fshirjes në çifte.

Metoda 5
Përdoren të gjitha n_c vëzhgimet e llogaritet ekuacionin e regresit të një variabli me të gjithë variablat e tjerë. Për shembull, le të jetë ky ekuacion regresi $X_1 = f(X_2, \cdots, X_p)$. Më pas nje vlerë e munguar e X_1 për rastin e j-të vlerësohet $\hat{x}_{1j} = f(x_{2j}, \cdots, x_{pj})$. Ekuacione të ngjashme mund të gjenden për $X_2, \cdots, X_p$. Tërësia e vrojtimeve që rezulton tashmë e plotësuar, përdoret për vlerësimin e $\overline{x}$ dhe S.

Metoda 6
Përdoret metoda 5 me ca ndryshime. Për shembull X_1 mund të parashikohet vetëm nga një variabël i vetëm $X_2, \cdots, X_p$ që është më i koreluar me X_1, ose X_1 mund të parashikohet nga nënbashkësia e $X_2, \cdots, X_p$. Metodat 5 dhe 6 quhen metodat e zëvendësimit të regresit.
Disavantazhi i parë i përdorimit të një prej këtyre metodave është që vetitë statistikore të tyre nuk njihen përveçse në raste të veçanta. Për më tepër këto metoda shpesh prodhojnë rezultate të shmangura. Rekomandimet tregojnë se vrojtimi dhe/ose variablat duhet të eleminohen, sipas rastit. Kështu vrojtimet dhe variablat duhet të eleminohen për të maksimizuar numrin e vrojtimeve të plota. Kështu, në se një vrojtim ka shumë variabla që mungojnë, duhet të eleminohet gjithë vrojtimi. Nga ana tjetër, në se një variabël mungon në shumë vrojtime, ky variabël eleminohet. Më pas përdoren procedurat e zakonshme të katrorëve më të vegjël ose procedurat shumëpërmasore.

2.6.2.1 Vërejtje
1. Shumica e paketave kompjuterike kanë si opsion' default' metodën e eleminimit të vrojtimit.
2. Në disa programe ekziston si opsion edhe fshirja në çift. Ky opsion mund të përdoret kur një vrojtues ka shumë variabla që i mungojnë pak vlera,dhe kur metoda e eleminimit të vrojtimit do të reduktojë numrin e vrojtimeve shumë. Përdoruesi duhet të ketë parasysh që mund të lindin dhe situata jo të vërteta në këtë rast nga llogaritjet

(si p.sh. shumë katrorësh ose vlera testesh F negative). Në rastet kur përdoret opsioni nuk aplikohet në mënyrë rigoroze teoria e nderthurjes statistikore.

2.7 Regresi i Shumëfishtë Stepwise

Regresioni i shumëfishtë shërben për të gjetur nënbashkësine më efektive të variablave përcaktues, në parashikimin e variablit të varur.
Variablat shtohen në ekuacionin e regresit duke përdorur kriterin e maksimizimit të R^2
Kur shtimi i variabave nuk sjell ndonjë ndryshim statistikor të rëndësishëm në R^2, atehere analiza e regresit të shumëfishtë Stepwise ndalon.

2.7.1 Shtrimi matematik i Regresi të Shumëfishtë Stepwise

Në shumë raste, gjatë aplikimit te regresit, vrojtuesi nuk ka informacion të mjaftueshëm rreth shkallës së rëndësisë së variablave të pavarur $X_1, X_2, \cdots, X_p$ në parashikimin e variablit të varur Y. Kontrolli i H_0: $\beta_i = 0$ për çdo variabël X_i, $i = 1, \cdots, p$, nuk na siguron këtë informacion.
Meqë statistika që përcakton 'efektivitetin' e një bashkësie variablash të pavarur në parashikim, është koefiçienti i korelacionit të shumëfishtë, një zgjidhje për problemin e mësipërm është llogaritja e koefiçientit të regresit të Y ndaj të gjitha nënbashkësive të mundshme të variablave të pavarur dhe, më pas, të zgjidhet nënbashkësia më e mirë, duke shfytëzuar këto koefiçientë.
Për çdo nënbashkësi me k, elementë, për $k = 1, \cdots, p$, përcaktohet nënbashkësia S_k që përmban koefiçientët më të mëdhenj të korrelacionit të shumëfishtë. Për regresin e Y ndaj nënbashkësisë S_1, kontrollohet hipoteza që përfshirja e $p - 1$ variablave të mbetur nuk përmirëson parashikimin e Y , me hipotezë alternative: verehet përmirësim. Për kontrollin e kësaj hipoteze përdoret testi statistikor i kontrollit te koefiçientëve 0, pra testi:

$$F = \frac{(SS_R{}' - SS_R)/m}{MS_R}$$

Nëse hipoteza hidhet poshtë, atëherë për regresin e Y në nënbashkësinë S_2, kontrollohet nëse përfshirja e $p - 2$ variablave të mbetur nuk përmirëson parashikimin e Y. Përsëritet testi në mënyrë të njëpasnjëshme, deri sa për nënbashkësite $S_m, 1 \leq m \leq p$, hipotezat që $p - m$ variablat e mbetur ,nuk përmirësojnë parashikimin e Y, pranohen. Atëherë S_m është nënbashkësia më e mirë e variablave në parashikimin e Y me qenë se: (a) ajo siguron koefiçientin ekorelacionit të shumëfishtë më të madh' ndërmjet të gjithë nënbashkësive me vellim m; dhe (b) përfshirja e $p - m$ variablave të mbetur nuk përmirëson ndjeshëm parashikimin e Y. Nëse kjo zgjidhje nuk është e vetme, natyra e problemit mund të sugjerojë nënbashkësinë më të përshtatshme.
Kur numri i variablave të pavarur është i madh, aplikimi I kësaj metode s'është dicka praktike, madje edhe në se kemi kompjuter me shpejtësi të madhe, përcaktimi i nënbashkësisë më të mirë duke përdorur këtë procedurë është I vështirë. Për shembull kur $p = 5$, atëherë janë 5+10+10+5+1=31 ekuacione regresi, dhe kur $p = 10$, janë 2(10+45+120+210)+252+1=1023 ekuacione regresi. Në përgjithësi, janë të nevojshëm $2^p - 1$ ekuacione regresi. Koha dhe shpenzimet e nevojshme na shtyjnë të gjejmë një rrugë tjetër për këtë problem.

Një zgjidhje është teknika e regresit hap pas hapi, në të cilën variablat e pavarur $X_1, X_2, \cdots, X_p$ futen njëri pas tjetrit në ekuacion, sipas disa kritereve të parapërcaktuara. Sapo një variabël futet në ekuacion, ai mund të ndërrohet me një variabël që është në modelin regresit, ose mund të hiqet fare nga ekuacioni. Bashkësia e kritereve që përcakton se si një variabël futet, ndërrohet ose hiqet quhet procedura stepwise. Diskutohen katër procedura stepwise.
Nga një procedurë stepwise, fitohet një listë e renditur parashikuesish. Për shembull në se $p = 5$ kjo listë mund të jetë$X_2, X_5, X_1, X_4\ dhe\ X_3$. Për të përcaktuar ekuacionin më të mirë nga kjo listë zgjidhen $m \leq p$ variablat e parë kështu që (i) ata parashikojnë Y sa më mirë që të jetë e mundur, dhe (ii) m është sa më e vogël që të jetë e mundur. Me fjalë të tjera një nënbashkësi variablash të varur zgjidhet nga lista e renditur, e tillë që të ketë aftësi parashikuese të lartë. Në shembullin e mësipërm mundet që nënbashkësitë të konsistojnë vetëm në $X_2\ dhe\ X_5$ dhe, se regresi me $X_2\ dhe\ X_5$është pothuaj aq i mirë sa regresi në$X_2, X_5, X_1, X_4\ dhe\ X_3$. Procedura për përcaktimin e m njihet më emrin procedura Iterative. Mëposhtë paraqiten tri rregulla të tilla ndërprerje.

2.7.2 Procedurat iterative

Le të supozojmë se kemi një bashkësi variablash të pavarur $X_1, X_2, \cdots, X_p$ të cilët janë të gjithë për tu shqyrtuar për parashikimin e Y. Le të supozojmë gjithashtu dhe që kemi një zgjedhje të rastit me vëllim n. Procedura ndëprerjes standarde, që do diskutohet fillimisht, konsiston në një rregull për të futur variablat dhe një rregull për heqjen e variablave (ndërrimi i variablave nuk përfshihet në procedurën e iterative standard). Siç do të shohim procedurat e tjera rrjedhin nga kjo procedurë.

1. Procedura standarde iterative (Metoda F).

Variablat futen ose hiqen duke përdorur një test statistikor, t testin, i cili është prezantuar për kontrollin e hipotezave, koefiçienti i korrelacionit të pjesshëm janë zero. Programet kompjuterike paraqesin katrorët e kesaj statistike, e cila ka shpërndarje F (me shkallë lirie të përcaktuar më poshtë).
Më pas supozojmë se një nënbashkësi c që përmban k variabla është futur në ekuacion, $k = 0,1, \cdots, p-1$. Më pas testi F teston nëse një variabël X (jo nga c) përmirëson në mënyrë të ndjeshme parashikimin e Y në krahasim me atë të variablave në c, p.sh. $H_0: \rho_{yx.c} = 0$, duke përdorur statistikën e testit F

$$F_{yx.c} = \frac{r_{yx.c}^2 (n-k-2)}{1-r_{yx.c}^2}$$

e cila nën hipotezën zero ka shpërndarje $F(1, n-k-2)$.
Në mënyrë të ngjashme merret edhe nje test statistikor po me shpërndarje Fisher, që teston procesin e heqjes, pra nëse një variabël X në një nëbashkësi c që përmban k variabla nuk nevojitet më. Pra nënbashkësia e zvogëluar c' që përmban $k' = k-1$ variabla, parashikon Y po aq mirë sa nënbashkësia c. Kështu, ne testojmë $H_0: \rho_{yx.c'} = 0$ duke përdorur testin statistikor me shpërndarje Fisher $F(1, n-k'-2)$:

$$F_{yx.c'} = \frac{r_{yx.c'}^2 (n-k'-2)}{1-r_{yx.c'}^2}$$

Siç do të shohim rregulli Iterativ sipas procedurës standard, përfshin zgjedhjen e një vlere minimale *F për pwrfshirjen e variablave* . Disa paketa programuese marrin një vlerë të paracaktuar 4.0. Për heqjen e variablave, zgjidhet gjithashtu një minimum i testit *F për të hequr* (më i vogël se minimumi *F për përfshirjen*; disa paketa programesh marrin minimumin F =3.9).
Hapat e procedurës standard më të detajuara janë:

Hapi 0:

Testohet $H_0: \rho_{yx_i} = 0, i = 1, \cdots, p.$ (Koefiçienti i korelacionit të thjeshtë është koefiçient i korelacionit të pjesshëm me $k = 0$ dhe c =bashkësi boshe.) Per këtë llogariten koefiçientët i korelacionit të thjeshtë r_{yx_i} dhe vlerat e F testit, F_{yx_i}, për $i = 1, \cdots, p$. Testi statistikor jepet nga

$$F_{yx_i} = \frac{r_{yx_i}^2 (n-2)}{1 - r_{yx_i}^2}$$

që përfitohet nga (*) me $k = 0$. Ai ka shperndarje Fisher, F me 1 dhe $n-2$ shkallë lirie.

Hapi 1

Variabli X_{i_1} qe ka vlerën e testit F më të madh, (ose ndryshe, katrorët e koefiçientëve të korrelacionit me Y më të mëdhenj) zgjidhet si parashikuesi më i mirë i Y. Nga ekuacioni i katrorëve më të vegjël,bëhet analiza e tabelës së variancës, dhe llogaritet koefiçienti i korrelacionit të shumëfishtë $r_{y.x_{i_1}} = \left| r_{yx_{i_1}} \right|$. Në këtë hap, testi F , që shërben për të hequr X_{i_1} ,është e njëjtë me vlerën e F-testit për të futur për X_{i_1}. Më pas llogariten koefiçientët e korelacionit të pjesshëm $r_{yx_i.x_{i_1}}$ dhe F-testet për përfshirjen e çdo njerit prej tyre

$$F_{yx_i.x_{i_1}} = \frac{r_{yx_i.x_{i_1}}^2 (n-3)}{1 - r_{yx_i.x_{i_1}}^2}$$

për $i = 1, \cdots, p, i \neq i_1$, për çdo variabël që nuk përfshihet në ekuacionin e regresit. Kjo statistikë F ka 1 dhe $n-3$ shkallë lirie, dhe kontollon hipotezat $H_0: \rho_{yx_i.x_{i_1}} = 0$ për $i = 1, \cdots, p, i \neq i_1$. Në se të gjitha F për të përfshirjen e variablave te tjerë janë më të vegjel se minimum i F për përfshirjen e variablit, atëherë ekzekutohet Hapi S. Në rast kundërt vazhdojmë me Hapin 2.

Hapi 2

Variabli X_{i_2} që ka vlerën e testit F për përfshirjen e variablit më të madh (që do të thotë, katrorin eientit të korrelacionit të pjesshëm më të madh me Y për X_{i_1} te dhënë) zgjidhet si parashikuesi më i mirë i Y nëse, më parë, ka qenë zgjedhur X_{i_1}. Llogariten nga ekuacioni i katrorëve më të vegjël, analiza e tabelës së variancave, dhe koefiçienti i korelacionit të shumëfishtë $r_{y.x_{i_1}x_{i_2}}$. Gjithashtu, llogariten $F_{yx_{i_1}.x_{i_2}}$ dhe $F_{yx_{i_2}.x_{i_1}}$. Këto statistika me 1 dhe $n-3$ shkallë lirie përcaktohen si më poshtë:

$F_{yx_{i_1}.x_{i_2}} = \frac{r^2_{yx_{i_1}.x_{i_2}}(n-3)}{1-r^2_{yx_{i_1}.x_{i_2}}}$ dhe $F_{yx_{i_2}.x_{i_1}} = \frac{r^2_{yx_{i_2}.x_{i_1}}(n-3)}{1-r^2_{yx_{i_2}.x_{i_1}}}$

Ato testojnë respektivisht hipotezat zero $H_0: \rho_{yx_{i_1}.x_{i_2}} = 0$ dhe $H_0: \rho_{yx_{i_2}.x_{i_1}} = 0$. Së fundmi, llogariten koefiçienti i korelacionit të pjesshëm $r_{yx_i.x_{i_1}x_{i_2}}$,dhe vlera e testit F me 1 dhe $n-4$ shkallë lirie për $i = 1, \cdots, p, i \neq i_1, i \neq i_2$për përfshirjen e x_{i_2}

$$F_{yx_i.x_{i_1}x_{i_2}} = \frac{r^2_{yx_i.x_{i_1}x_{i_2}}(n-4)}{1-r^2_{yx_i.x_{i_1}x_{i_2}}}$$

të cilat bëjnë të mundur kontrollin e hipotezës $H_0: \rho_{yx_i.x_{i_1}x_{i_2}} = 0$. Në se të gjitha F për përfshirjen e x_{i_2} janë më të vogla se se minimum i vlerës se statistikës F për përfshirjen x_{i_1}, atëherë ekzekutohet Hapi S. Në të kundërt, kalohet në Hapin 3.

Hapi 3

(a) Le të shënojmë me L bashkësinë e l variablave të pavarur që janë përfshirë në ekuacionin e regresit. Nëse ndonjë nga vlerat e statistikës F për të hequr variablin e përfshirë në L, është më I vogel se minimum i paracaktuar i F për të hequr, atëherë variabli me F -vlere për të hequr më të vogël nuk përfshihet më në L, dhe ekzekutohet hapi (3b) ku l zëvendësohet me $l-1$. Në se të gjitha vlerat e F testit për të përfshirë variablat që tashmë s'janë në L, janë më të vogla se minimum i F të paracaktuar të përfshirjes, atëherë ekzekutohet Hapi S. Në të kundërt variabli me vlerën e F përfshirjes më të madh merret dhe, i shtohet L-së kështu që l zëvendësohet me $l+1$.

(b) llogariten për Y dhe variablat në L, ekuacioni i katrorëve më të vegjël, tabela e analizës së variancave dhe koefiçienti i korrelimit të shumëfishtër$r_{y.l}$. Gjithashtu llogaritet vlerat e testeve F për të hequr parametra, $F_{yx_{i_j}.(l-1)}$ ndërmjet Y dhe X_{i_j} në L në se $l-1$ variabla mbesin në L. Kjo ka 1 dhe $n-l-1$ shkallë lirie dhe testohet hipoteza $H_0: \rho_{yx_i.(l-1)} = 0$. Së fundmi llogariten koefiçienti i korrelacionit të pjesshëm $r_{yx_i.l}$dhe testi F për perfshirjen $F_{yx_i.l}$ për Y dhe X_i jo në L për variabla të dhënë në L. Kjo statistikë ka 1 dhe $n-l-2$ shkallë lirie dhe testohet $H_0: \rho_{yx_i.l} = 0$ për X_i jo L, $i = 1, \cdots, p$.

Hapat 4,5,…

Përsëritet Hapi 3 në mënyrë në çdo iteracion. Hapi S ekzekutohet (a) kur vlerat e testit F për përpërfshirjen e variablave të , që s'janë në L, janë më të vogla se minimum i statistikes F për përfshirjen e variablit, (b) kur të gjithe variablat janë përfshirë dhe vlera e testit F për të hequr për variablat e përfshirë është më e madhe se F për heqjen e variablit, vlerë kjo e paracaktuar në fillim, ose (c) kur numri i variablave të përfshire në regresin e shumëfishtë është $n-2$.

Hapi S

Zakonisht, me kërkesë të përdoruesit, jepet një tabelë përmbledhëse. Për çdo hap afishohen: numri i hapit, variabli i perfshirë apo i hequr, vlera e F testeve për perfshirjen dhe heqjen e variablit në fjalë, dhe koefiçienti i korrelacionit të shumëfishtë të Y me variablat e përfshirë në këtë model.

2. Procedura Standarde me Zëvendësim (FSWAP)

Kjo procedurë përdor të njëjtat rregulla për heqjen dhe shtimin e variablave si procedura e mësipërme iterative, përveç që në çdo hap, mund të bëhen ndërrime ndërmjet variablave në ekuacion. Kjo procedurë është si të thuash më e mira, sipas variantit që dëshirojmë, dhe ekuacionit të regresit të marrë sipas metodës më sipërme. Duke përdorur këtë procedurë, një variabël në ekuacion ndërrohet me një variabël që nuk është në ekuacion, nëse ndërrimi rrit koefiçientin e korelacionit të shumëfishtë (madje edhe nëse nuk siguron ndonjë rritje të ndjeshme). Në një hap të caktuar, nëse janë k variabla në nënbashkësinë c të cilët janë përfshirë në ekuacion, procedura iterative (i) heq variabla nga c duke përdorur testin F për heqje, (ii) ndërrimin e një variabli në c me një variabël jo në c, dhe (iii) shtimin e një variabli në c duke përdorur testin F përfshirës.

3. Metoda e Korelacionit të Shumëfishtë (Metoda R)
Kjo procedurë përdor të njëjtin rregull për përfshirjen e variablave në ekuacionin e shumëfishtë, por perdor një rregull tjetër për heqjen e variablave. Në një iteracion të caktuar, ajo përdor rregullin R^2, sipas të cilit një variabël hiqet nga modeli i regresit linear, në se siguron një katror të koefiçientit të korelacionit të shumëfishtë më i madh. Është e mundur të arrihet rritje e R^2, meqenëse R^2 nuk është vetëm një funksion i madhësive të përcaktuara (siç janë, n dhe ${s_y}^2$), por edhe të dy madhesive që ndryshojnë (s^2 mesatares së katrorit të mbetjes dhe p numri i variablave në ekuacion). Kështu është e mundur për këto sasi të ndryshojnë,në atë mënyrë që një variabël më pak të japë një R^2 më të madhe. Kjo procedurë përfundimisht konsiston sipas rradhës (i) heqjen e një variabli, duke përdorur rregullin R^2, dhe (ii) shtimin e një variabli, duke përdorur testin F të përfshirjes.

4. Metoda e Korelacionit të Shumëfishtë me Ndërrim (RSWAP)
Kjo procedurë është e njëjtë me metodën R, por përdor edhe rregullni i ndërrimit. Rendi i kësaj procedure është: (i) heqja e variablit duke përdorur rregullin R^2, (ii) ndërrimi i dy variablave për të rritur R^2, dhe (iii) shtimi i një variabli duke përdorur testin F të përfshirjes.

2.7.2.1 Përfundime
Shumica e paketave të programeve të regresit stepwise ka një opsion për ndërhyrjen në procesin e gjetjes së variablave më me interes në ekuacionin e regresit. Për çdo variabël përdoruesi specifikon, me anë të një parametri që fut, *le ta quajmë nivel detyrimit*, një instruksion që përcakton nëse variabli do të futet pavarësisht nga vlera e testit F të përfshirjes. Kështu përdoruesi ka kontroll mbi variablat (në kundërshtim me zgjedhjen statistikore përshkruar më sipër), dhe në këtë mënyrë variablat që kanë vlerë më të madhe në parashikim mund të futen të parët. Procedura stepwise përfshin më pas variablat që kanë një nivel rëndësie të vogël.

2.7.3 Rregullat e Ndërprerjes

Do paraqesim tri rregulla për përcaktimin e numrit të variablave (predictors) për gjetjen e ekuacionit më të mirë të regresit. Rregulli standard, i cili është i pranishëm në shumicën e paketave të programeve statistikore, kontrollon numrin e variablave që futen në ekuacion me anë të një parametri që përcaktohet nga përdoruesi dhe, që

quhet *minimum vlerë e testit F të përfshirjes*. Siç thamë më sipër ky parametër është njëvlerë minimum e statistikes Fisher që i korespondon nivelit të rëndësisë α, $F_{1-\alpha}(1, v)$. α rekomandohet të jetë 0.15, me gjithë se shumë përdorues vendosin α në 0.05.

1. Rregulli Standard i Ndërprerjes

Rregulli Standard I Ndërprerjes për të fituar bashkësinë më të mirë të parashikuesve H mund përcaktohet lehtë nga tabela përmbledhëse e printuar në Hapin S. Në një modelin iterativ, vlera e testit F të përfshirjes për çdo variabël të futur krahasohet me vlerën e F për përfshirje të përcaktuar më parë.
Më të detajuara hapat ne procedurat kompjuterike janë:
(a) Në hapin 1, futet një variabël X_{i_1}. Nëse vlera e testit F përfshirës është jo e konsiderueshme, d.m.th. F përfshirjes<*minimum i F të perfshirjes se paracaktuar*, atëherë përfshirja e variablit në fjale në regres është e pa kuptimtë, dhe përdoruesi duhet të kërkojë metoda të tjera për të analizuar të dhënat. Në të kundërt, $H = \{X_{i_1}\}$
(b) Në Hapin 2, futet një variabël X_{i_2}. Nëse vlera e testit F të përfshirjes së tij<*minimum i F të përfshirjes së paracaktuar*, atëherë H mbetet vetëm me variablin X_{i_1}, dhe regresi më i mirë është ai i Hapi 1. Në të kundërt, $H = \{X_{i_1}, X_{i_2}\}$.
(c) Për çdo hap të mëtejshëm, në se futet një variabël, krahasohet vlera e testit F përfshirës me *minimumin e F përfshirës*. Në se F përfshirëse e variablit është e konsiderueshme, shtohet variabli në H dhe shkohet në hapin tjetër. Nëse jo, ndërpritet shtimi i H dhe merret, si ekuacioni më i mirë i regresit, ai i marrë në hapin e mëparshëm.

2. Rregulli i Ndërprerjes Bazuar në Ndryshimet në R^2

Kjo procedurë kërkon që përdoruesi të paracaktojë me kujdes vlerën *'minimale të testit F të përfshirjes* ' dhe vleren*'minimale të F për të hequr variablin e përfshirë*'. Vlera *'minimale e testit F të përfshirjes* ' mund të zgjidhet në mënyrë të tillë që çdo variabli që parashikohet të jetë i nevojshëm në parashikimin t'I japim shanse 50-50. Kështu, vlera *'minimale e testit F përfshirjes* ' duhet të jetë $F_{0.50}(1, n-p-1)$. Në anën tjetër,vlera *'minimale e testit F për të hequr variablin e përfshirë*' mund të zgjidhet në mënyrë të tillë që, të kemi shanse të vogla që një variabël i përfshirë, të hiqet. Kështu zgjidhet vlera *'minimale të F për të hequr variablin e përfshirë*' shumë i vogël, p.sh. 0.01. Më pas vëmendja përqëndrohet tek variablat që janë në tabelën përmbledhëse në hapin e fundit. Le të jetë L bashkësia e l variablave, $l \leq p$ dhe, le të jetë $r_{y.l}$ koefiçienti i korelacionit të shumëfishtë ndërmjet Y dhe variablave në L. (Nëse një apo më shumë variabla hiqen, është e nevojshme të llogaritet serisht $r_{y.l}$). Le të jetë H bashkësia e h variablave në ekuacionin e regresit në një hap të ndërmjetëm. Procedura bën testimin e hipotezës $H_0: \rho_{y.h} = \rho_{y.l}$ duke përdorur

$$F = \frac{n-l-1}{l-h}\frac{r_{y.l}^2 - r_{y.h}^2}{1-r_{y.l}^2}$$

Nëse H_0 është e vërtetë, testi F ka shpërndarje $Fisher$ me $l-h$ dhe $n-l-1$ shkallë lirie. Kryhet ky test në mënyrë të njëpasnjëshme, për çdo iteracion derisa fitohet vlera e parë, e testit F josignifikative. Supozojmë se kjo ndodh p.sh. në

Iteracion 3, ku Y është shprehur me një ekuacion regresi të një bashkësi H të h variablave. Më pas në se në Iteracion 4 duke dashur të përfshijmë variabla te rinj në ekuacion regresi, nesë vlreat e F testit jane josignifikative atehere ndalohet së shtuari H-në dhe marrim si ekuacionin më të mirë të regresit, atë të fituar në Iteracion 3. Nëse testi që kryhet synon të heq variablat ekzistues, atëhere variabli hiqet, kryejmë kontrollin e mësipërm për Iteracion 4. Në se testi përfshirës është signifikativ, atëherë H është bashkësia e variablave e fituar në Iteracion 4, dhe marrim si regresin më të mirë atë të Iteracion 4. Në se vlera e F testit është josignifikative, përcaktojmë H me $h-1$ variabla si bashkësinë e variablave që marrin pjesë në këtë ekuacion regresi.

3. Rregulli i Ndërprerjes Bazuar në Mesataren e Katrorit të Gabimit të Pakushtëzuar

Bendel dhe Afifi[6] (1976) rekomandojnë një rregull ndalimi alternativ. Ata propozojnë të kontrollohet hipotezat që gabimi i pakushtëzuar i katrorit mesatar (UMSE)[7] nuk ulet nga një hap në tjetrin. Kjo madhësi përcaktohet si $UMSE = E(Y-\hat{Y})^2$, ku vlera e parashikuar e Y, $\hat{Y}$, fitohet nga tërësia e vlerave $Y, X_1, \cdots, X_p$ të cilët mendohet se kanë shpërndarje normale shumëpërmasore. Vlerësimi i UMSE në një hap të dhënë behet nga statistika:

$\widehat{UMSE}(q,n) = \frac{n^2-n-2}{n(n-q-2)}\, MS_R = \frac{(n-1)(n^2-n-2)(1-r^2_{y.c})s_y{}^2}{n(n-q-1)(n-q-2)}$,

ku $q$është numri i variablave në ekuacion dhe MS_Rështë mbetja e katrorit mesatar në hapin e dhënë.

2.8 Zbatim 2

Në punimin me teme: **Cfarë ndikon në memorje**, ku synohet të vlerësohet ndikimi ne memorjen e adoleshentëve të disa variablave si, mënyra e të ngrënit, konsumimi apo jo i alkolit, sasia e orëve të fiskulturës, (aktiviteteve sportive) dhe gjinia. Përcaktues për memorjen, në studimin ne fjalë, është marrë variabli 'nrwords', që përcakton numrin e fjalëve që mund të memorizojë një adolishent pas leximit të 40 fjaleve për 5 minuta. Studimi i ndikimit të 5 variablave ne fjale, në variablin tone, u mendua të behej me ane të modelit të regresit linear të shumëfishtë. Por për ndertimin e këtij modeli, një ndër kushtet që duhet të plotësohet është dhe ai që, variablat që marrin pjesë në model të jenë dy e nga dy të pavarur. Për të kontrolluar këtë u shkruajt dhe u ekzekutua kodi i mëposhtëm në SPSS :

```
PARTIAL CORR
  /VARIABLES=sportlevel duhanpojo alkolpojo gjinia BY nrtotalIfjaleve
  /SIGNIFICANCE=TWOTAIL
  /STATISTICS=DESCRIPTIVES
  /MISSING=LISTWISE.
```

Afishimet paraqiten në tabelen 2-5.

Tabela 2–5 Tabela e koefiçientëve të korelacionit

Correlations

Control Variables			sportlevel	duhan po/jo	alkol po/jo	gjinia f/m
nr total I fjaleve	sportlevel	Correlation	1.000	.090	.091	.272
		Significance (2-tailed)	.	.121	.116	.000
		df	0	295	295	295
	duhan po/jo	Correlation	.090	1.000	.444	.325
		Significance (2-tailed)	.121	.	.000	.000
		df	295	0	295	295
	alkol po/jo	Correlation	.091	.444	1.000	.263
		Significance (2-tailed)	.116	.000	.	.000
		df	295	295	0	295
	gjinia f/m	Correlation	.272	.325	.263	1.000
		Significance (2-tailed)	.000	.000	.000	.
		df	295	295	295	0

Koefiçientat e korelacinit të afishuara mes çdo dy variablave në shqyrtim tregojnë se çdo dy prej variablave"*alkol','duhan',e 'sport level'*, mund të konsiderohen të pavarur linearisht. Meqënëse ky kusht plotësohet mund të flasim për mundësine e ndërtimit të nje modeli regresi me këta variabla. Ndersa vërehet që ka njefare korelacioni të dobët të gjinise me variablat e tjerë në shqyrtim.

2.9 Regresi i Shumëfishtë Hirarkik

Metoda në fjalë shihet në programet kompjuterike si regresi i shumëfishtë hirarkik
Ajo përdoret zakonisht në rastet, kur gjatë studimit diskriptiv të të dhenave, vërehet varesi e variablit në studim nga një variabel tjetër i pavarur, por ne nuk jemi të interesuar aq shumë për këtë varësi. Interesi ynë është përqëndruar më shumë në kontrollin e varësisë së variablit në studim nga variabla të tjerë.

Në regresin e shumëfishtë hirarkik futja e variablave të pavarur bëhet në dy faza.
Në fazën e parë futen variablat e pavarur që ne duhet t'i kontrollojmë më parë në lidhjen e regresit. Ndersa ne fazën e dytë futen ato variabla të pavarur që ne duhet t'i shqyrtojmë, pasi varësia e parë lineare merret parasysh.
Vleresimi i ndikimit të variablave, me të cilen punohet në fazën e dytë, bëhet me anë të testit statistikor që ndertohet duke pasur parasysh R^2 e marrë nga faza e pare.

2.9.1 Zbatim 1(vazhdim)

Nëse i referohemi sërisht zbatimit **"Efekti I SO_2 në prodhimet bujqësore"**, vihet re perdorimi i kësaj metode studimi të të dhenave. Në ketë zbatim u vu re se ndotja industriale ndikon ne prodhueshmërine e produkteve bujqesore jonxhes, patates, perimes dhe fasules.
Mesatarja e prodhimit te jonxhes është statistikisht më e lartë ne zonën e ndotur, derisa diferenca e studiuar e vleres se komponentit në fjalë në zonën e ndotur nga zona e kontrollit është negative.
Mesatarja e patates, perimes dhe fasules është statistikisht më e ulët në zonën e ndotur.
Por, sasia e produkteve bujqësore dihet që mund të varet nga kushtet abiotike të lartëpërmendura. Për të testuar nëse këto tri variablat e lartë përmendura ndikojnë në parashikimin e sasisë së prodhimit te *patates, fasules, perimeve* dhe *jonxhes*, është pëdorur metoda e regresit hirarkik të shumëfishtë[(5)].
Si variabli i pavarur i listuar më parë, the independent variable listed first, u përdor variabli zone, dhe si variabla të shtuar , the added variables ,distance, ujrat e zeza dhe mbetjet urbane. Përgjithsisht interesi nuk qëndron tek varësia e variablave të kontrollit'*area*' ndaj variablave të varura, ndaj jemi në kushtet e përdorimit të regresit hirarkik.
Afishimet paraqesin rezultatet e t-testeve,nese gjate ndertimit të hiperplaneve të regresit nuk I marrim parasysh faktoret abiotik, dhe tregojnë effekt të faktorit *'sistem i ujrave të zeza'*

Tabela 2–6 Afishimet e t-testeve dhe koefiçienteve te regresit

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-301.111	1619.402		-.186	.853
	area	2405.556	915.482	.208	2.628	.009
2	(Constant)	-9952.987	7017.577		-1.418	.158
	area	5846.296	2787.910	.505	2.097	.038
	sistemi i ujrave te zeza	267.888	1363.239	.016	.197	.044
	terheqja e mbeturinave	1375.448	521.163	.248	2.639	.009
	distanca shtepi-pusne meter	3.503	4.381	.180	.800	.425

a. Dependent Variable: jonxhe

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	229.364	22.768		10.074	.000
	area	-59.364	12.871	-.349	-4.612	.000
2	(Constant)	-14.914	98.412		-.152	.880
	area	43.286	39.097	.255	1.107	.270
	sistemi i ujrave te zeza	5.775	19.118	.023	.302	.063
	terheqja e mbeturinave	10.801	7.309	.133	1.478	.142
	distanca shtepi-pusne meter	.164	.061	.572	2.662	.009

a. Dependent Variable: perime

Faktoret abiotik sistemi i ujrave të zeza dhe, ne menyre jo shumë te sigurt dhe ai i mbetjeve urbabe, kanë ndikim ne produktivitetin e bimeve. Sa më shumë kujdes të tregohet në to nga ana sanitare aq më pozitivisht ato ndikojne tek bimët.
Informacioni i marrë nga variablat e shtuar, redukton parashikimin në ndryshueshmërinë e prodhueshmerisë se bimëve, vetëm si rezultat i zonës, me 20% per perimet, 4.3% for për jonxhën.

2.9.2 Zbatim 2:(vazhdim)

Në punimin me teme: **Cfarë ndikon në memorje**, pas analizës deskriptive të të dhënave është vënë re ndikim në variablin '*nrwords*,' dhe në përmiresimin e tij për sigurimin e kushtit të normalitetit 'sqrtnrwords', ndikim i variablit '*gjini*a', ndikim i cili për ne nuk paraqet interes. Ndaj per të percaktuar efektin apo jo ne variablin në shqyrtim të kater variablave të tjerë te pavarur, që marrin pjesë në sdtudimin në fjale, efektin e duhanit alkolit te të ngrënit shëndetshëm dhe të sportit është kryer regresi linear hirarkik. Kjo metodë synon, pas heqjes nga *sqrtword* e efektit të gjinise, të kontrollojë për efekt të alkolit, duhanit të të ngrenit shëndetshëm dhe të sportit.

Tabela 2–7 Tabela e regresit hirarkik

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.186[a]	.035	.031	.67475	.035	10.583	1	296	.001
2	.647[b]	.418	.408	.52735	.384	48.150	4	292	.000

a. Predictors: (Constant), gjinia f/m

b. Predictors: (Constant), gjinia f/m, hani shendetshem po/jo, alkol po/jo, sa here bejne sport ne jave, duhan po/jo

Tek tabela 2-7 vlerësimi i R^2-change tregon se, ky model justifikon 39% ndryshimet ne rrënjen katrore të numrit te fjalëve të memorizuara. Kontolli i hipotezës R=0, pra nuk ka lidhje lineare mes modelit te propozuar dhe faktoreve te tjere te testuar bie poshte. Kjo, sepse vlera e testit te kontrollit të hipotezës në fjalë është signifikativ, vlera e testit F jep p=0.000002. Madje R=0.612 tregon se lidhja ne fjale, e shprehur nga modeli i regresit, është e fortë.

Tabela 2–8 Koefiçientët e korelacionit të regresit hirarkik

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.335	.118		28.184	.000
	gjinia f/m	.260	.080	.186	3.253	.001
2	(Constant)	3.267	.108		30.143	.000
	gjinia f/m	.068	.070	.049	.972	.332
	sa here bejne sport ne jave	.223	.018	.597	12.470	.000
	duhan po/jo	-.148	.075	-.104	-1.972	.050
	alkol po/jo	-.146	.072	-.104	-2.035	.043
	hani shendetshem po/jo	.042	.070	.028	.607	.544

a. Dependent Variable: sqrtnword

Kryerja e metodes se regresit linear hirarkik konkludon për efekt të sportit, duhan dhe alkolit (lexo p- vlerat në tabelen 2.8). Ndërsa variabli i mënyrës së të ngrënit, nuk sjell ndonjë ndryshim nëse nuk merret parasysh në modelin e regrest linear.
Koefiçientet negativ pranë variablit duhan e alkol tregojnë efektin negativ te dy këtyre faktorëve, dhe efekt pozitiv të sportit në memorje.

Për të përcaktuar se cilët nga variablat e pavarur në fjalë luajnë rol me të madh në sqrtnrwords kontollohen vlerat e koefiçientëve të korelacionit të pjesshëm, tabela 2.9 Vlerat tregojnë për ndikim të sportit ne masen 0.6^2=0.36, pra 36% e ndryshimeve ne variablin e shqyrtuar shpjegohen nga variabli *'sa herë bëjnë sport në javë'*. Ndërsa ndryshimet në duhan dhe alkol shpjegojnë vetëm 4% të ndryshimeve të '*sqrtnrword*'

Tabela 2–9 Koefiçientët e korelacionit të pjesshëm

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	sa here bejne sport ne jave	.629[b]	13.085	.000	.606	.897
	duhan po/jo	-.240[b]	-4.141	.000	-.234	.921
	alkol po/jo	-.224[b]	-3.906	.000	-.222	.950
	hani shendetshem po/jo	.072[b]	1.251	.212	.073	.993

a. Dependent Variable: sqrtnword
b. Predictors in the Model: (Constant), gjinia f/m

2.9.3 Zbatim 3 (vazhdim)

Në shembullin e mësipërm për të parë se cili nga 4 variablat e pavarura sporti, alkoli, te ushqyerit dhe duhani, ndikojnë më shumë në parashikimin e 'sqrtnrword', është përdorur regresi linear stepwise.
Afishimet e mëposhtme tregojnë se cilët jane variablat kryesorë që ndikojnë në këtë model regresi.

Tabela 2–10 Regresi stepwise

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	saherebejnesportnejave	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

2	duhan po/jo	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: sqrtnword

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.624[a]	.389	.387	.53684
2	.639[b]	.409	.405	.52887

a. Predictors: (Constant), saherebejnesportnejave

b. Predictors: (Constant), saherebejnesportnejave, duhan po/jo

Metoda e regresit stepwise tregon se nga tere faktoret e konsideruar te pavarur ndikim tek *srtnword* ka *numri i oreve te fiskultures* dhe *konsumojne apo jo duhan*.
Faktori *kosumim alkoli, mënyra e të ushqyerit* janë eleminuar nga teresia e faktoreve te pavarur te futur ne shqyrtim. Modeli justifikon rreth 40% te ndryshimeve te variablit sqrtnword. Madje vetëm variabli sport justifikon 38.96% të ndryshimeve në sqrtwords.

Nese shikohet koefiçienti para tij ne modelin e regresit te shumefishte linear stepwise, eshte 0.23 modeli i ndertuar vetem ne baze te faktorit sa here bejne sport ne jave kalimi nga dy here fiskulture ne jave ne tri here fiskulture do te ne jave do te sjell mesatarisht 13.69 nga 11.7 fjale të memorizuara.

3 Analiza e Variancës Dypërmasore (Metoda ANOVA Dypërmasore)

3.1 Modeli me nenpopullime te fiksuara *(Modeli I)*

Me këtë metode, vëzhgohet ndikimi i dy faktorëve të ndryshëm ***A*** e ***B***, mbi një variabë, Y. Kjo metodë i shikon nënpopullimet të përcaktuara nga nivelet e secilit prej dy faktorëve, si dhe nga nderthurja e tyre. Faktorët A e B, në fakt janë variabla të tipit kategorik.Vlerat e ketyre dy variablave te pavarur percaktojne nivelet e faktoreve në shqyrtim dhe, kombinimet e nivele të ndryshme të tyre, përcaktojnë modelin që po shqyrtojmë. Kështu, nëse A ka *I* nivele(variabli A ka I vlera) dhe B ka *J* nivele, (variabli B ka J vlera), atëhere, *IJ* janë kombinimet e niveleve të dy faktorëve. Kombinimi i nivelit të *i* –të të faktorit A, me atë të *j*-të të faktorit B përcaktojnë atë që mund të quhet *qeliza ij*.

Në Analizën Dypërmasore dallohen rastet kur numri i vëzhgimeve të variablit të rastit Y, për çdo qelizë të jetë i njëjtë dhe kur ai s'është I tillë. Këto vëzhgimë përcaktojnë zgjedhjen e nënpopullimit të përcaktuar nga ndërthurja e nivelit të i-të te A-së, me nivelin e j-te të B-së.

Ndaj ndërtohet modeli dy përmasor, ku çdo vlerë e n.r. në shqyrtim paraqitet si kombinim i mesatares, i nivelit të ndikimit të faktorëve në shqyrtim, te nderthurjes se tyre dhe të një termi gabimi.

Shënojmë y_{ijk} vlerën e *k*-të të Y në qelizën e *ij*-të, dhe ndërtohet modeli për shqyrtimin e efekteve të dy faktorëve të fiksuar

$$Y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+e_{ijk}$$

ku i=1..I; j=1..J; k=1..K

μ është mesatarja, α_i eshtë efekti i faktorit A, dhe β_j efekti i faktorit B, $(\alpha\beta)_{ij}$ paraqet ndërthurjen e dy faktorëve ij-të . Gabimet e rastit e_{ijk} janë variabla të rastit të pavarur, dhe kanë shperndarje $N(0, \sigma^2_e)$.

Në rastin kur $(\alpha\beta)_{ij}=0$ për çdo *i* dhe *j* , modelet A dhe B quhen aditive.

Le të kalojmë tani në vlerësimin e parametrave që përcaktojnë efektin e faktorëve në shqyrtim.

Vleresimet e këtyre parametrave me Metodën e Katrorëve më të Vegjël nuk janë të vetëm, ndaj vihen edhe disa kushte të tjera. Supozohet se :

$$\sum_{i=1}^{I} \alpha_i = 0$$

$$\sum_{j=1}^{J} \beta_j = 0$$

$$\sum_{i=1}^{I}(\alpha\beta)_{ij}=0 \quad \text{për j=1..J} \qquad\qquad \sum_{j=1}^{J}(\alpha\beta)_{ij}=0 \quad \text{për i=1..I}$$

Si zakonisht, duke përdorur Metodën e Katrorëve më të Vegjël, përcaktohen vlerësime të pazhvendosura të parametrave të këtij modeli, μ, β_j, α_i. Shuma e marrë është kjo:

$$S=\sum_i\sum_j(y_{ijk}-\mu-\alpha_i-\beta_j-(\alpha\beta)_{ij})^2$$

dhe, gjejmë vlerësuesit e parametrave:

$$\hat{\mu}=\bar{y}$$
$$\hat{\alpha}_i=\bar{y}_{i..}-\bar{y}$$
$$\hat{\beta}_j=\bar{y}_{i..}-\bar{y}$$
$$(\alpha\beta)_{ij}=\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}$$
$$\text{për i=1..I dhe j= 1..J}$$

duke vënë në dukje se $\bar{y}$ është mesatarja e vlerave të $y_{ijk;}$

$\bar{y}_{.j.}$ është mesatarja e kollonës së j-të

$\bar{y}_{i..}$ është mesatarja e rreshtit

Nga teorema e Gaus Markovit këto vlerësime të pavarura të parametrave sigurojnë një minimum të variancës nga të gjithë vlerësimet që fitohen si kombinim linear i vlerave të vrojtuara. Kështu shuma *S* është bazë për vlerësimin e pazhvendosur të $\sigma^2_{\epsilon.}$ Kjo shumë, mund të ndahet në katër shuma, sipas burimeve të variacionit, A,B,AB dhe Mb(Regresit).

$$SKT=\sum_i\sum_j\sum_k(y_{ijk}-\bar{y})^2=JK\sum_i(\bar{y}_{i..}-\bar{y})^2+IK\sum_j(\bar{y}_{.j.}-\bar{y})^2+K\sum_j\sum_i(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y})^2+$$
$$\sum_i\sum_j\sum_k(y_{ijk}-\bar{y}_{ij.})^2=SK_A+SK_B+SK_{AB}+SK_R$$

Anova Dy Përmasore shqyrton katër komponentët e variancës[9], duke ndertuar statistika te pazhvendosura per çdo burim te variacionit, si tabela 3.1.

Për të përcaktuar ndryshimin midis grupeve sipas një faktori vlerësojmë shumën e katrorëve të gabimeve sipas secilit faktor S_{KA} dhe S_{KB} dhe kështu mesatarja e gabimit për çdo faktor do të jetë $M_A=\frac{S_{KA}}{I-1}$ dhe $M_B=\frac{S_{KB}}{J-1}$

Për të kuptuar nëse ka ndryshime midis grupeve të ndryshme sipas secilit prej dy faktorëve, duhet që ato të krahasohen me variancën e gabimit $б^2$

Tabela 3–1 Përbërësit e variacionit Modeli 1

Burimet e variacionit	Shuma e katrorëve të gabimeve	Mesatarja e katroreve
Faktori A	$S_{KA}=JK\sum_i(\bar{y}_{i..}-\bar{y})^2$	$M_{KA}=\frac{S_{KA}}{I-1}$
Faktori B	$S_{KB}=IK\sum_j(\bar{y}_{.j.}-\bar{y})^2$	$M_{KB}=\frac{S_{KB}}{J-1}$
Ndërthurja e dy faktorëve	$S_{KAB}=K\sum_j\sum_i(\overline{y_{ij.}}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2$	$M_{KAB}=\frac{S_{KAB}}{(I-1)(J-1)}$
Mbetjet	$S_{KR}=\sum_i\sum_j\sum_k(y_{ijk}-\bar{y}_{ij.})^2$	$M_{KR}=\frac{S_{KR}}{IJ(K-1)}$
Totali	$S_{KT}=\sum_i\sum_j\sum_k(y_{ijk}-\bar{y})^2$	

.

Ajo supozohet e njëjtë për të gjithë grupet. Ndaj, për vlerësimin e saj përdoret statistika

$\hat{\sigma}^2=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}K_{ij}s^2{}_{ij.}}{IJ\sum(K_{ij}-1)}=\frac{\sum\sum\sum(y_{ijk}-\bar{y}_{ij})^2}{IJ\sum(K_{ij}-1)}\equiv M_{KR}$ që na siguron një vlerësim të panjohur të σ^2

Meqë y_{ijk} janë një zgjedhje nga një popullim i madh më shpërndarjë $N(\mu, \sigma^2)$ atëherë $\frac{(IJK-1)S_{KT}}{\sigma^2}\sim X^2(IJ(K-1))$

Ndërtohen testet statistikore më shpërndarje të përcaktuar ,Tabela3.2, dhe në bazë të tyre kontrollon testet statistike të paraqitura në Tabelën 3.2.

Tabela 3–2 Testet sipas hipotezave

Ho: $(\alpha\beta)_{ij}=0$	Ho: $\alpha_i=0$	Ho: $\beta_j=0$
$F=\frac{M_{KAB}}{M_{KR}}$	$F=\frac{M_{KA}}{M_{KR}}$	$F=\frac{M_{KB}}{M_{KR}}$
((I-1)(J-1);IJ(K-1))	((I-1);IJ(K-1))	((J-1);IJ(K-1))

Në bazë të p-vlerave, pra të sipërfaqeve në të djathtë nën kurbat e këtyre shpërndarjeve, diskutohen :

Në fillim hipoteza Ho: $(\alpha\beta)_{ij}=0$,

- Nëse p-vlera është shumë, shumë e vogël, atehere nuk kemi ndonjë ndërthurje midis dy faktorëve.

- *Vazhdohet*, duke shfrytezuar p-vlerat e testeve perkatës, për t'i dhënë përgjigje pyetjeve nëse ka efekt secili prej dy faktoreve A dhe B. Modeli I i Anova-s i jep pergjigje pyetjes, a jane të njejta mesataret e nenpopullimeve te caktuara sipas secilit faktor, apo jo.
- Nese p-vlera është e madhe , atëhere shumës së katrorëve të gabimeve të mbetjeve, duhet ti shtojmë dhe këtë shumë tashmë të konsiderueshme të ndërthurjes midis dy niveleve, duke marrë kështu një vlerësim tjetër të σ^2: $M_{SP}=S_{KP/}\ \upsilon_P$ ku $S_{KP}=S_{KR}+S_{KAB}$ me një numër shkallësh lirie $\upsilon_P=\upsilon_{R+}\ \upsilon_{AB[]}$.
- Nderhyrja në sintaksën e programeve statistikore bën të mundur dhe kontrollin e hipotezës për ndërthurjen apo jo të dy faktorëve në studim.

3.1.1 Përfundime

Duhet vënë në dukje se nuk ka ndonjë vlerë kufi të përcaktuar të p-vlerës, për të thënë se kur duhet bërë dhe kur jo ky modifikim, çdo gjë varet nga personi që studion problemin.
Megjithate ky modifikim keshillohet te bëhet në rastet kur p-vlera prane nderthurjes se dy faktoreve, P>0.5. Mosmarrja parasysh e këtij konkluzioni, mund të cojë në përfundime të gabuara në analizën e dy testeve të tjera.
Punohet më pas për shqyrtimin e ndikimit apo jo një ë nga një të faktorëve A dhe më pas B B, duke shfrytëzuar përkatësisht $F=M_{SA}/M_{SP}$ për testimin e hipotezës Ho: $\alpha_i=0$ dhe $F=M_{SB}/M_{SP}$ të hipotezës Ho: $\beta_j=0$.

3.1.2 Zbatim 1

Në një studim, "**Cilët faktorë ndikojne në memorje**", është kontrolluar me parë ndikimi i shumë faktorëve që mendohej të ndikonin në ndryshoren e rastit, që përcakton sasinë e fjalëve të memorizuara, (n.r. që për të plotësuar kushtin e normalitetit është modifikuar në *sqrtnrword*). Faktori *gjini* si dhe *numri i orëve që merren me sport ne jave*, janë përcaktuar të dy si faktorë ndikues në *sqrtnrword*.
Për të vrojtuar si ndikojnë këto dy faktore, në përputhje dhe me problemin qe trajton studimi, ndikimin tek memorja e adoleshentëve, është bërë krijimi i një variabli te ri kategorik, *nivsportlevel* (faktor me J=3 nivele). Me metodën e ANOVA-s Dypërmasore, modeli me nënpopullime të fiksuara, është kontrolluar ndikimi i faktorit *gjini* (I=2 nivele) dhe *sportlevel.* Tabela *3.3* paraqit të dhënat e kësaj analize.

Afishimet e para të tabelës 3.3 përmban përvec mesatareve në çdo 6 kombinime të dy faktorëve të përcaktuar dhe vellimin e zgjedhjes për çdo 2*3=6 qelizat e krijuara.
Ndërsa, në pjesën tjetër të tabelës 3.3 lexojmë përfundimet e kësaj metode. Vlera e statistikës F=.103 dhe e p-vlerës koresponduese .902, tregojnë nuk mund të flasim për nderthurje mes faktoreve[10]. Madje dhe vlera e variancës qe përcakton kjo ndërthurje është tej mase e vogël , siç lexohet në tabelë vetëm 0.01.
Ndërsa afishimet prane faktoreve tregojnë, për ndikim *sportlevel* (F=42.256 dhe e p-vlerës 7.62 E-17), dhe për mos ndikim të faktorit *gjini*, (F=.333 dhe p-vlera .565).

Tabela 3–3 Tabela e të dhënave deskriptive dhe analizës ANOVA dypërmasore

Descriptive Statistics
Dependent Variable: sqrtnword

gjinia f/m	sportlevel	Mean	Std. Deviation	N
femer	0-2 here sport	3.4465	.65658	137
	3- 4 here sport	4.0243	.59823	35
	me shume se 4 here sport	4.3390	.60750	7
	Total	3.5944	.69653	179
mashkull	0-2 here sport	3.4594	.47177	59
	3- 4 here sport	4.1089	.52193	34
	me shume se 4 here sport	4.4163	.52522	26
	Total	3.8540	.64051	119
Total	0-2 here sport	3.4504	.60572	196
	3- 4 here sport	4.0660	.55943	69
	me shume se 4 here sport	4.3999	.53454	33
	Total	3.6981	.68555	298

Dependent Variable: sqrtnword

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	37.779[a]	5	7.556	21.672	.000	.271
Intercept	2148.261	1	2148.261	6161.623	.000	.955
gjinia	.116	1	.116	.333	.565	.001
sportlevel	29.465	2	14.733	42.256	.0000000	.224
gjinia * sportlevel	.072	2	.036	.103	.902	.001
Error	101.806	292	.349			
Total	4215.000	298				
Corrected Total	139.585	297				

a. R Squared = .271 (Adjusted R Squared = .258)

3.2 **Modeli me nënpopullime të rastesishme *(Modeli II)***

Tabela Anova-s nund te shfrytezohet edhe ne një kendveshtrim tjetër, sipas metodes me nenpopullime te rastesishme[].
Keshtu, popullimi $N(\mu,б^2)$ ndahet ne nenpopullime te percaktuara sipas dy faktoreve ne shqyrtim, si rezultatat i kombinimit të tyre.
Zgjedhja e marre siguron, zgjedhje per çdo nepopullim me shperndarje $N(m_{ij}, б^2_{ij})$ ku i=1..I; j=1..J.
Nga zgjedhjet e ketyre grupeve, llogariten me pare mesataret m_{ij} per çdo grup, te caktuar nga kombinimi i dy faktoreve te shqyrtuar A dhe B.
Vlerat e zgjedhjes fillestare (si zgjedhje e nenpopullimit te grupit *ij*) mund te shkruhen ne formen:
$y_{ijk}= m_{ij}+e_{ijk}$ ku e_{ijk} te pavarura me shperndarje $N(0, б^2)$
(Ne fakt vellimi i zgjedhjes per çdo grup eshte i ndryshem, pra k ndryshojne, keshtu simbolika me e pershtatshme per k do te ishte k_{ij}. Por, nese paraqes rastin kur vellimi ne çdo grup te percaktuar nga dy faktoret eshte i njejte, nuk kam ndonje ndryshim, vetem na thjeshton simboliken)
Nenpopullimet me shperndarje $N(m_{ij},\sigma^2_{ij})$, sherbejne tani si zgjedhje per popullimin e madh, te pare si popullim i nenpopullimeve. Ndaj elementet e kesaj zgjedhje shkruhen ne formen:
$m_{ij}=\mu+a_{i.}+b_{.j}+(ab)_{ij}$,
ku $a_i \sim N(0,\sigma^2_{i.})$ te pavarura, $b_{.j} \sim N(0,\sigma^2_{.j})$ te pavarura dhe te pavarura edhe $(ab)_{ij} \sim N(0,\sigma^2_{ij})$

Perfundimisht, ndertohet modeli $Y_{ijk}=\mu+a_i+b_j+ (ab)_{ij}+e_{ijk}$ ku i=1..I; j=1..J; k=1..K, ku μ mesatarja e gjithe popullimit, a_i, b_j dhe $(ab)_{ij}$ jane te pavarura dhe kane shperndarje me mesatare 0 e variance perkatesisht σ_a, σ_b, σ_{ab} dhe $e_{ijk} \sim N(0, \sigma^2+\sigma^2_{ij})$.
Shtrohet problemi i vleresimit te ketyre parametrave. Duke perdorur te njejten metode vleresohet

$$\hat{\sigma}_a = (M_{KA}-M_{KAB})/JK,\ \hat{\sigma}_b=(M_{KB}-M_{KAB})/IK \text{ dhe } \hat{\sigma}_{ab}=(M_{KAB}-M_{KR})/K$$

Per te sqaruar ku ndryshojne dy metodat ne ndarjen ne katër komponente te variancës, eshte paraqitur tabela 3.4.

Tabela 3–4 Përbërësit e variacionit Modeli II

Burimet e variacionit	Perberesit sipas modelit I	Perberesit sipas modelit II
Faktori A	$\sigma^2+\frac{JK\sum \alpha_i^2}{I-1}$	$\sigma^2+K\sigma^2_{ab}+JK\sigma^2_a$
Faktori B	$\sigma^2+\frac{IK\sum \beta_j^2}{j-1}$	$\sigma^2+K\sigma^2_{ab}+IK\sigma^2_b$
Ndërthurja e dy faktorëve	$\sigma^2+\frac{JK\sum\sum (\alpha\beta)_{ij}^2}{(I-1)(J-1)}$	$\sigma^2+K\sigma^2_{ab}$
Mbetjet	σ^2	σ^2

Ndaj testet statistikore sipas modelit II do te jene:

Ho: $(\sigma_{ab})^2=0$	Ho: $\sigma_a{}^2=0$	Ho: $\sigma_b=0$
F= $\frac{M_{KAB}}{M_{KR}}$	F= $\frac{M_{KA}}{M_{KAB}}$	F= $\frac{M_{KB}}{M_{KAB}}$
((I-1)(J-1);IJ(K-1))	((I-1);(I-1)(J-1))	((J-1);(I-1)(J-1))

Në fillim kontrollohet hipoteza Ho: $(\alpha\beta)_{ij}=0$,

- Nëse p-vlera është shumë, shumë e vogël, atehere nuk kemi ndonjë ndërthurje midis dy faktorëve.

Modeli II i Anova-s *vazhdon*, duke shfrytezuar p-vlerat e testeve perkatës, dhe i jep pergjigje pyetjes, a ka apo jo variacion midis nenpopullimeve te caktuara sipas secilit faktor.

- Nese p-vlera është e madhe, atëhere shumës së katrorëve të gabimeve të mbetjeve, duhet ti shtojmë dhe këtë shumë tashmë të konsiderueshme të ndërthurjes midis dy niveleve, duke marrë kështu një vlerësim tjetër të σ^2: $M_{SP}=S_{KP/}\upsilon_P$ ku $S_{KP}=S_{KR}+S_{KAB}$ me një numër shkallësh lirie $\upsilon_P=\upsilon_{R+}\upsilon_{AB}$.

Vazhdohet me shqyrtimin e ndikimit apo jo të faktorëve A dhe B, duke shfrytëzuar përkatësisht $F=M_{SA}/M_{SP}$ për testimin e Ho: $\alpha_i=0$ dhe $F=M_{SB}/M_{SP}$ të Ho: $\beta_j=0$.

Shihet se statistikat F ndryshojne nga ato te modelit te pare, e kjo ne rastet e nje nderthurje te dy faktoreve me njeri tjetrin, na ben qe vleresimi i marre me modelin II te ANOVA-s te jete ndoshta ndryshe nga ai sipas modelit I.

3.2.1 Zbatim 2

Diskutimi i problemit të ndikimit të ndotjes në pneumologji me Anoven Dy Përmasore

Për problemin pneumologjik kemi zgjedhur si tregues përcaktues ndryshorjen e rastit nivkoll
Në studimin e problemit të pneumologjisë, është vëne re se faktori zonë ka ndikuar në nivelin e kollës dhe, është shtruar problemi i ndikimit apo jo mbi këtë ndryshore rasti të kohëqëndrimit në zonën industriale (kohëqëndrim, faktori A me 4 nivele)[8] dhe të shkallës së duhanpirjes (nivduhan, faktori B me tri nivele).
Ky problem u trajtua me metodën e Anova-s dy përmasore.
Tabela e Anovës dy permasore shfaq F= 3.54 dhe P= 0.039 pranë variablit *kohëqëndrim*, që dëshmon se roli i kohqëndrimit në këto zona, per ndryshoren nivkoll është domethënës.
Ndersa prane variablit shkalla e duhan pirjes p-vlera =0.056, qe deshmon per mosndikim te duhanit ne ndryshoren e rastit ne shqyrtim.
Sipas *Modelit II* , modeli i ANOVA-s Dypermasore të rastwsishme, marrim këto vlerësime për përbërësit e variancës, Tabela 3-5:

Tabela 3–5 Afishimet e përbërësve të variacionit

Përbërësi i variacionit	Vlera
niveli kohëqëndrimit	0.10873
niveli i duhan pirjes	0.08524
niveli kohëqëndrimit* niveli i duhan pirjes	-0.06260
mbetjet(gabimet)	0.50283

Meqë varianca e një komponenti nuk mund të jetë më e vogel se 0, atëhere vleresimi per variancën nën efektin e ndërthurjes së dy faktorëve merret 0.
Pra, jemi ne rastin kur dy faktoret konsiderohen aditive, dhe kjo metode për shqyrtimin e variancës se dy faktorëve afishon prane faktorit nivelkoheqendrim F= 8.93dhe P=0.043 dhe prane faktorit nivel i duhanpirjes F= 7.44 dhe P=0.026, dhe tregojnë tani **ndikimin e dy faktorëve në shqyrtim në ndryshoren nivkollë**. Kështu meqe faktorët nuk ndërthuren, ne zgjidhjen e ketij problemi mund te jipet edhe përgjigja se të dy faktorët ndikojnë në n.r në shqyrtim, pra në këtë tregues të problemeve pneumologjike.
Pra, mund të themi se fakti që modelin e ANOVA-s e shikojmë sipas **modelit I apo II, na con në rezultate tërësisht të ndryshme.**

Por, kontrolli i hipotezës së ndërthurjes apo jo të dy faktorëve në shqyrtim shfaq, pranë vlerës se statistikës F= 2.3367, p vlerën 0.077 , tregon se ndonese nuk ka nje nderthurje te qarte mes dy faktoreve, keto faktore *nuk mund te konsiderohen si aditiv*. Kjo na bën të mos ndikohemi shume nga vlerat e afishuara, por të nisemi nga vlerësimi tjetër . Ne baze te sqarimeve te bera , modifikohen vlerat e emeruesve te statistikave.
Fillimisht modifikohet shuma $S_{KP}=S_{KR}+S_{KAB}$ = 59.83 dhe me pas statistika ne emeruesin e dy statistikave vleresuese: $M_{SP}=S_{KP/}\ \upsilon_P$ =0.4913 që shfaq F=2.3367 dhe P=0.089.
Përfundimisht, një vëzhgim më i kujdesshëm i Anovas Dypërmasore, tregon se në bazë të kësaj metode nuk mund të themi se ndikimi i kohëqëndrimit në këtë zonë industriale është domethënës mbi këtë tregues të problemeve pneumologjike.
Po kështu modifikohet dhe vlerësimi për ndikimin apo jo të faktorit tjetër, shkallës së konsumimit të duhanit. Duke u nisur nga vlerësimi i mësipërm i M_{KP} , merret F=1.729 dhe P=0.18, qe do te thote se nuk rezulton ndikim vetem i ketij faktori ne ndryshoren e rastit qe shqyrtohet.
Nëse do ti rishikonim rrezultatet e ti grumbullonim ato, duke përfillur ndërthurjen mes dy grupeve atëhere, do të fitonim të njëjtat F-vlera si në rastin e parë, pra të njëjtat përgjigje për ndikimin e faktorëve në shqyrtim.
Pra, shtojmë se sa të ndryshme mund të jenë përfundimet varet, dhe nga qëndrimi që mbahet ndaj ndërthurjes se dy faktorëve.

3.2.2 Vërejtje

SPSS shoqërohet me tri editor, editori i të dhënave, i output-ve dhe editori i sintaksës.

Editori i të dhënave përmban variablat, llojin e tyre e bën të mundur kryerjen e analizës së ndryshme statistikore.
Editori vier i outputeve, krijon lehtësi për afishimin e tabelave nga komanda të ndryshme të ekzekutuara nga editori i të dhënave dhe ai i sintaksës.
Editori i sintaksës hapet File/Data/Editor/New/ Syntax, dhe bën të mundur ruajtjen e sintaksës së gatshme për bashkësi të dhënash të ndryshme. Kështu editori i vecante mund të ekzekutohet në bazën e të dhënave të reja, duke ndryshuar vetëm emertimet e variablave.
Afishimet për përcaktimin e ndërthurjes mes dy faktorëve arrihen pasi në këtë editor të fundit shkruhet kodi i posaçëm.

3.3 Shtrimi matematik i Analizës së Kovariancës (ANCOVA)

Bazë e analizës së kovariancës është gërshetimi i metodës ANOVA me një analizë të thjeshtë të regresit linear.
Kuptohet, se ajo përdoret në rastet kur ndryshorja e rastit. nuk varet vetëm nga nivelet e një faktori, por dhe nga një ndryshore rasti tjetër, e vazhdueshme.
Pra, supozohet se kemi p.sh. I nivele të një popullimi sipas një faktori. Shënohen y_{ij} vëzhgimet e marra në eksperimentin e j-të
të nivelit të i-të. Supozohet gjithashtu se, y_{ij} kanë shpërndarje $N(\mu_i, \sigma^2)$.
Në të tilla raste modeli i ndërtuar matematik gjatë analizës ANOVA është $y_{ij}= \mu+ \alpha_i+ e_{ij}$
ku e_{ij} kanë shpërndarje $N(0, \sigma^2)$
μ pritja matematike e popullimit,
α_i ndikimi i nivelit të i-të dhe, siç e dihet, për të siguruar një vlerësim të këtyre parametrave me metodën e katrorëve më të vegjël, vihet kushti $\sum_i J_i \alpha_i = 0$

Por, ndërkohë krahas vrojtimeve te vlerave y_{ij} në vëzhgim e i-të të nivelit të j-të, bëhet një tjetër vëzhgim i variablit x, që supozohet të jetë linearisht i varur me y , vëzhgime që shënohen me $x_{ij.}$
Vlerësimi sipas ANOVA-s do të bëhej më i saktë nëse y_{ij} do ti zbritej ajo pjesë që përcaktohet nga regresi linear me x_{ij}.
Kështu nga vlerat e marra y_{ij} , fitohen vlera të reja $y_{ij}^*=y_{ij}-\beta(x_{ij}-\bar{x})$
Këto, quhen vlera të reduktuar, sepse i është hequr efekti i ndikimit tek to i x_{ij}, duke u supozuar ky një ndikim linear.
Modeli matematik ANOVA, tani ndërtohet për këto vlera të reduktuara y^*_{ij}.

$$y^*_{ij} = \mu+ \alpha_i+ e_{ij} \text{ ku :}$$

e_{ij} kanë shpërndarje $N(0, \sigma^2)$
μ pritja matematike e popullimit të madh,
α_i ndikimi i nivelit të i-të.
Ky model matematik i vlerave të reduktuar nuk është gjë tjetër vetëm se modeli matematik :

$$y_{ij} = \mu+ \beta(x_{ij}-\bar{x})+ \alpha_i+ e_{ij}$$

i vlerave të y_{ij} , që mund ta quajmë dhe model të analizës së kovariancës.

Përfundimisht, në këtë model, $\beta(x_{ij}-\bar{x})$ është ndikimi në y_{ij} i variablit x, i supozuar si një ndikim linear dhe α_i, ndikimi i nivelit të i-të në sajë të faktorit në shqyrtim.
Kështu, meqë modeli i analizës së kovariancës kthehet në një model të përgjithshëm linear, duke përdorur metodën e katrorëve më të vegjël, përcaktohen vlerësime të pazhvendosura të parametrave te ketij modeli, μ, β, α_i.
Pas ndertimit të shumës $S = \sum_i \sum_j (y_{ij} - \beta(x_{ij} - \bar{x}) - \mu - \alpha_i)^2$
kërkohet minimumi i saj, duke përcaktuar kështu këto vlerësime për parametrat:

$$\frac{\partial S}{\partial \mu} = -\sum_i \sum_j (y_{ij} - \beta(x_{ij} - \bar{x}) - \mu - \alpha_i) = 0$$

$$-n\bar{y} - n\mu = 0$$

$$\hat{\mu} = \overline{y}$$

$$\frac{\partial S}{\partial \beta} = -\sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \beta(x_{ij} - \bar{x}) - \mu - \alpha_i) = 0$$

$$\sum_i \sum_j (x_{ij} - \bar{x}) y_{ij} - \beta(x_{ij} - \bar{x})^2 - (\mu + \alpha_i)(x_{ij} - \bar{x}) = 0$$

$$nga \cdots ku$$

$$\hat{\beta} = \frac{\sum_i (x_{ij} - \bar{x}_i)(y_{ij} - \overline{y_i})}{\sum_{i,j} (x_{ij} - \overline{x_i})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\frac{\partial S}{\partial \alpha} = -\sum \sum_j (y_{ij} - \beta(x_{ij} - \bar{x}) - \mu - \alpha_i) = 0$$

$$\hat{\alpha}_i = \frac{\sum_i \sum_j y_{ij}}{n} - \hat{\beta} \sum_i \sum_j (x_{ij} - \bar{x})$$

$$\hat{\alpha}_i + \hat{\mu} = \bar{y}_i - \hat{\beta}(\bar{x}_i - \bar{x})$$

Dihet që këto vlerësime të parametrave të modelit linear të ndërtuar, sigurojnë dhe një minimum të variancës nga të gjithë vleresuesit linear të pazhvendosur të këtyre parametrave.
Ndaj,

$$S = \sum_i \sum_j (y_{ij} - \hat{\beta}(x_{ij} - \bar{x}) - \hat{\mu} - \hat{\alpha}_i)^2 = \sum_i \sum_j (y_{ij} - \frac{S_{xy}}{S_{xx}}(x_{ij} - \bar{x}) - \bar{y}_i - \frac{S_{xy}}{S_{xx}}(\bar{x}_i - \bar{x}))^2 =$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_i - \frac{S_{xy}}{S_{xx}}(x_{ij} - \bar{x} - \bar{x}_i + \bar{x}))^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i - \frac{S_{xy}}{S_{xx}}(x_{ij} - \bar{x}_i))^2 = S_{yy} - \frac{S^2_{xy}}{S_{xx}}$$

është një vlerësim i zhvendosur i variancës së gabimit.

Ndërsa, $MS = \frac{1}{n - I - 1}(S_{yy} - \frac{S^2_{xy}}{S_{xx}})$ është vlerësim i pazhvendosur për σ^2

Në këtë mënyrë është ndërtuar një vijë e regresionit linear

$$\hat{y} = \bar{y}_i + \frac{S_{xy}}{S_{xx}}(x - \bar{x}_i) \quad \text{për i=1..I.}$$

Për këtë arsye, këto quhen vija të regresionit linear brenda grupit.
Pas dhënies së komandës së vlerësimit me ANOVA të modelit të modifikuar do të afishohen këto vlera:

Tabela 3–6 ANOVA e modelit të qendërzuar

Midis niveleve	$SS_M = \sum_i J(\overline{y^*_i} - \overline{y^*})^2$	$SS_M/(I-1)$
Brenda niveleve	$SS_B = \sum_i \sum_j (y^*_{ij} - \overline{y^*_i})^2$	$SS_B/(n-I-1)$
Totali	$SS_T = \sum_i \sum_j (y^*_{ij} - \overline{y^*})^2$	$SS_T/(n-1)$

Kështu në rastin tonë, tabela 3-6 paraqet vlerat e analizës së variancës për vlerat e modifikuara. Ndërsa, për vlerat reale të vrojtuara, nuk është gjë tjetër vecse një analizë kovariance.

Tabela 3–7 Përbërësit e analizës së kovariancës mes grupeve

Midis niveleve	$SS_M = \sum_i J(\overline{x_i} - \bar{x})((\overline{y_i} - \bar{y})$	$SS_M/(I-1)$
Brenda niveleve	$SS_B = \sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \overline{y_i})$	$SS_B/(n-I-1)$
Totali	$SS_T = \sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y})$	$SS_T/(n-1)$

4 Analiza e Variancës Shumëpërmasore MANOVA

Kjo metodë është si të thuash një metodë ANOVA, por në rastin kur kemi shumë variabla të varur. Nga pikpamja e përgjithshme ANOVA teston për diferencën midis mesatareve të dy ose më shumë grupeve, në këtë këndveshtrim metoda MANOVA teston për diferencën në dy ose me shume mesatare vektorësh. Duke qenë një problem i komplikuar, problemi në fjalë, për zgjidhjen e tij jane dhenë mënyra të ndryshme. Në këtë kapitull do të paraqesim disa variante për zgjidhjen e tij, dhe do të diskutohet për prioritetet e secilës metodë, duke konkluduar kështu dhe për rastet kur secila nga metodat e trajtuara do të ishte mirë të përdorej.

4.1 Analiza e Variances Shumëpërmasore (MANOVA)

Shtrimi teorik i metodës MANOVA përcakton te veçantën e kësaj metode. Kjo metodë bazohet në një model regresi linear ndaj shpesh njihet me emrin MANOVA e variantit GLM.
Nga n individë, pra nga një zgjedhje me vellim n, janë vëzhguar vlerat e p variablave të rastit, dhe vëzhgimet janë shënuar si më poshtë:

Variablat	Individët					
	1	2	.	.	.	n
1	Y_{11}	Y_{12}				Y_{1n}
.	.	.	.	.	.	.
.	.	.			.	.
.	.	.			.	.
p	Y_{p1}	Y_{p2}	.	.	.	Y_{pn}

Pra, vektori $Yi=(y_{i1},y_{i2},\ldots y_{in})$ paraqet vlerat koresponduese te *n* vëzhgimeve në tërë individet të variablit të i-të. Modeli linear për çdo komponente $\mathbf{Y_i}$, i=1,2,…,p, që mund të shihet si nje model me nje variabël, përcaktohet nga barazimi vektorial:

$$E(\mathbf{Y_i})=\mathbf{X'}\boldsymbol{\beta_i}, \quad \text{cov}(\mathbf{Y_i})= \sigma_{ij}\,\mathbf{I}$$

ku $(\mathbf{X'})^{n*m}$ është matrica e desinjuar e rankut $r \leq m < n$, σ_{ii} është dispersioni I variablit të i-të dhe
$\boldsymbol{\beta_i}= (\beta_{i1},\ldots, \beta_{im})'$ është vektori me m komponentë, komponentët e të cilit janë parametra të panjohur të variablit të i-të.
Varesia e variablave shprehet nga

$$\text{cov}(\mathbf{Y_i},\mathbf{Y_j}) = \sigma_{ij}\,\mathbf{I} \qquad i,j = 1,2\ldots.,p$$

ku σ_{ij} është kovarianca mes variablit të i-të dhe të j-të. Është supozuar se $p \leq n-r$ dhe $m<n$
Modeli mund të shkruhet ne formë matricore

$$Y=X'\beta\_+e$$

ku vlerat e vrojtuara përcaktojnë matricën

$$\begin{bmatrix} y_{11} & y_{21} & . & y_{p1} \\ y_{12} & y_{22} & . & y_{p2} \\ . & . & . & . \\ . & . & . & . \\ y_{1n} & y_{2n} & . & y_{pn} \end{bmatrix}$$

(**X'**) ^{n*m} është matrica e desinjuar e rankut *r*, dhe

$$\begin{bmatrix} \beta_{11} & \beta_{21} & . & \beta_{p1} \\ \beta_{12} & \beta_{22} & . & \beta_{p2} \\ . & . & . & . \\ . & . & . & . \\ \beta_{1n} & \beta_{2n} & . & \beta_{pn} \end{bmatrix}$$

është matrica e parametrave të panjohur.
Së fundmi, e $^{n}*^{p}$ është matricë, rreshtat e së cilës janë një zgjedhje e rastit n përmasorë që formohet nga shperndarja p- përmasore **N(0,Σ)** ku $\mathbf{\Sigma}^{p*p}$ është matrica e kovariancës dhe $\mathbf{0}^{p*1}$ është vektori 0.
Ekuacioni I marë nuk është gjë tjetër veç se mënyra formale e shprehjes se modelit të regresit të shumëfishtë linear.

4.1.1 Zbatimi 1

Është përdorur metoda MANOVA, Modeli i Regresit Linear të saj (GLM), në trajtimin e **Ndikimit të nivelit te ndotjes ne problemet shendetësore të përgjithshme, ne problemet dermatologjike dhe ato pneumatike.** Kështu, në problemin në fjalë jane marë të dhëna nga 171 individë, banues te tri zonave, dy të konsideruara industriale, Marinëz dhe Sheqishtë-Belinë, dhe, nje e konsideruar e paster Kurjan. Të intervistuarve i jane berë dhe 13 pyetje te ndryshme, për gjendjen shëndetesore të të intervistuarve.
Në këtë situatë n=171, p=13 dhe Y_i është vektori i 171 vëzhgimeve, përgjigjeve të pyetjes së i-të, për i=1,2,…,13. Për këtë (pyetje), variabel të varur të i-të, modeli do të shkruhej në formën:

$$E(Y_i)=X'\beta_i$$

pra $Yij=\mu_i*1+\alpha_{i1}X_{j1}+ \alpha_{i2}X_{j2}+e_{ij}$, për i=1,2,…,13 dhe j=1,2,…,171.
Kështu $\beta i=(\mu_i, \alpha_{i1}, \alpha_{i2})'$ është vektori me 3 parametra i faktorit të i-të. X_{j1} dhe X_{jn} janë indekset e variablit '*fshatrat*'
Ishte statisticieni David P. Nichols[(12)], që më 1997, pasi bëri një analizë të dallimeve mes MANOVA-n dhe Metoden e Regresit Linear te Shumefishtë të një variabli kategorik, tregoi se Metoda e Regresit Linear te Shumefishtë nuk siguron një zgjidhje të vetme të këtij ekuacion matricor. Një ndër zgjidhjet më të thjeshta mund të kërkohet duke u bazuar në nje matrice njësi. Ndërsa në MANOVA, modeli e Regresit Linear te Shumefishtë kërkohet të jetë kanonik, ndaj modeli I matrices X merr formën si ne tabelën 4-1. Kjo, sepse kollona e fundit duhet shprehur si kombinim linear i kollonave para saj, pra A3 = C - A1 - A2

Tabela 4–1 Modeli i Matrices sipas GLM dhe MANOVA-s

	Matrica GLM				Matrica e MANOVA-s		
Nivelet	C	A1	A2	A3	C	A1	A2
1	1	1	0	0	1	1	0
2	1	0	1	0	1	0	1
3	1	0	0	1	1	-1	-1

Në rastin tonë m= r =3 modeli mund të shkruhet në këtë formë matricore:

$$\begin{bmatrix} y_{11} & y_{21} & \cdot & y_{13,1} \\ y_{12} & y_{22} & \cdot & y_{13,2} \\ . & . & . & . \\ . & . & . & . \\ y_{1,63} & y_{2,63} & \cdot & y_{13,63} \\ y_{1,64} & \cdot & \cdot & y_{13,64} \\ . & . & . & . \\ y_{1,116} & \cdot & \cdot & y_{13,116} \\ y_{1,117} & \cdot & \cdot & y_{13,117} \\ . & . & . & . \\ y_{1,1712} & \cdot & \cdot & y_{13,171} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ . & . & . \\ . & . & . \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ . & . & . \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ . & . & . \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 & \mu_2 & \cdot & \mu_{13} \\ \alpha_{1,1} & \alpha_{2,1} & \cdot & \alpha_{13,1} \\ \alpha_{1,2} & \alpha_{2,2} & \cdot & \alpha_{13,2} \end{bmatrix} + e$$

4.1.2 Vlerësimi i parametrave

Vleresusesi për matricen e parametrave të panjohur përcaktohet me vlerësimet që merren me Metoden e Katroreve më të Vegjel, duke shfrytëzuar vlerat e vëzhguara Y_i pra të variablit të i-[të].

Kështu zgjidhja e ekuacionit

$(XX')\ \beta_i = XY_i$

na siguron vlerësuesit që na duhen të $\beta_i = (\beta_{i1},\ldots, \beta_{ip})$, e që mund të shihen si nje model linear një permasor për Y_i.

Vlerësuesi i pazhvendosur për σ_{ij} nga teoria e modelit me një variabël përfitohet si më poshtë:

$$\frac{R_0(i,j)}{n-r} = \frac{Y_i'Y_i - (Y_i'X')\widehat{\beta_i}}{n-r}$$

ku r = rank **X'.** Vlerësuesi I pazhvendosur për σ_i është

$$\frac{R_0(i,i)}{n-r} = \frac{Y_i'Y_i - (Y_i'X')\widehat{\beta_i}}{n-r}$$

për i,j= 1,2,…,p. $R_0(i,j)$ quhen shumat e katroreve të mbetjeve dhe $R_0(i,j)$ është shuma e prodhimit të mbetjeve i,j = 1,2,…,p . Matrica

$$\begin{bmatrix} R_0(1,1) & R_0(1,2) & . & R_0(1,p) \\ R_0(2,1) & R_0(2,2) & . & R_0(2,p) \\ . & . & . & . \\ . & . & . & . \\ R_0(p,1) & R_0(p,2) & . & R_0(p,p) \end{bmatrix}$$

quhet matrica e shumës se katroreve të mbetjeve.

4.1.3 Testi I Hipotezës se Linearitetit

Testi i hipotezës së Modelit të Regresit Linear është trajtuar në 1965 nga Rao. Ai, për secilin nga variablat e varur, kryente ANOVA-n një-përmasore. Kështu ai analizonte hipotezën :

$$\text{Ho: } H'\beta_i=\xi_i \quad \text{për} \quad i=1,2,..p$$

Këto hipoteza në trajtë matricore mund të shkruhen në formën

$$H'\beta=\{\xi_1\ \xi_2\ ,\ldots\xi_p]$$

ku matrica $\mathbf{H'}^{s^*\ m}$, me rank s≤k dhe, vektori $\xi_i^{\ s^*1}$ janë të dhënë. Kështu kemi një model të përcaktuar, nga i cili mund të sigurojme nje vlerësim per β_i, βi^*, dhe matricë produkt dhe shumë katrorësh të mbetjeve $\mathbf{R_1}$. Matrica $\mathbf{R_1\text{-}Ro}$ quhet matrica produkt dhe shumë katrorësh që përcaktuan nga devijimin nga hipoteza. Analiza e variances, duke patur parasysh zberthimin e **R1** = **Ro** +$\mathbf{R_1\text{-}Ro}$ kthehet, në problemin e krahasimit të **Ro** me $\mathbf{R_1\text{-}Ro.}$

Më 1965 Rao tregoi se hipoteza e mësipërme e trajtës matricore merr formën e gjetjes së *p* rrenjëve $\lambda_1, \lambda_{2,\ldots}\ \lambda_p$ të ekuacionit

$$|R_0 - \lambda R_1|=0$$

Testi I kontollit te hipotezës zero kthehet kështu në teste të funksioneve te λ_i.

Një mënyrë është të gjesh minimumin e rrënjëve, sepse kështu arrihet maksimumi i devijimeve nga hipoteza zero, dhe teston për të.

Ndersa sipas kriterit Willks kriteri Λ përcaktohet nga

$$\Lambda=\lambda_1\ \lambda_{2\ldots}\ \lambda_p=\left|\frac{R_0}{R_1}\right|$$

4.1.4 Zbatimi 1 (*vazhdim*)

Në rastin tone testohet përmes metodes ANOVA njëpërmasore hipotezat H_0 marrin formën:

$$\text{Ho: } \alpha_{i1}=\ \alpha_{i2}=0$$

për i=1,2,…,13.

4.1.5 Testi i diferences mes vlerës se mesatareve midis disa popullimeve:MANOVA

Le të shënojmë me $n_1,n_2,\ldots n_k$ vëzhgimet e kryera për *k* popullime, dhe të shenojme me $\overline{Y_{ir}}$ mesataret e zgjedhjeve të *p* variablave nga zgjedhja e *r*-të, për *r* = 1,2,…,k.

Ndërsa $S_{ij}^{(r)}$elementi I matricës së shumës së katrorëve me n_r-1 shkallë lirie që merret nga zgjedhja e r-të..
Dhe së fundmi, le të jenë $\overline{Y_1}, \overline{Y_2}, \ldots \overline{Y_p}$ vlerat mesatare dhe, (S_{ij}) matrica e shumës se katrorëve të fituar nga kombinimi I të gjithë zgjedhjeve në një të vetëm. Duke vepruar si tek ANOVA një përmasore përcaktojmë:
$B_{ij} = \sum_1^r n_r \overline{Y_{ir}}\ \overline{Y_{jr}} - \overline{Y_i}\ \overline{Y_r}$

si shumë të produkteve midis popullimeve dhe
$\ddot{E}_{ij} \sum_{r=1}^{k} S_{ij}^{(r)}$

shumën e produkteve brenda popullimit të i-të , për i=1,2,…,p. Këto madhësi afishohen në tabelën e MANOVA-s . Kriteri Λ, në te cilen kjo metodë bazohet për të vlerësuar barazimin e vlerave të mesatareve është:

$\Lambda = \frac{|W|}{|W+B|}$
ku $|W|$dhe $|W+B|$janë përcaktorët e matricave (W_{ij}) dhe (B_{ij})+ (W_{ij}). Statistika Λ ka shpërndarje të vështirësi në përcaktim,ndaj në praktikë përdoret përafrimi *F*:

$$F = \left(\frac{ms - 2\lambda}{2r}\right)\left(\frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}}\right)$$

me $m = n - 1 - \frac{1}{2}(p + k)$ $\qquad s = \sqrt{\frac{p^2(k-1)^2 - 4}{p^2 + (k-1)^2 - 5}}$

Kështu për përcaktimin e përqindjeve përdoret shpërndarja F(2r,(ms-2λ)) shkallë lirie .
Idea e përafrimit te shpërndarjes së vërtetë me shpërndarjen e Fisherit është bërë nga Rao, dhe testi në fjalë njihet me emrin e tij. Por ky test përdoret atëhere kur variabli I pavarur që studjohet, plotëson kushtin e normalitetit dhe, natyrisht, ate te Homogjenintetit te Variancës , që kontrollohet me anë të testit Leven.
Ndërsa Bartlett e përafroi këtë shpërndarje me shpërndarjen X^2.
Sipas Bartlett transformimi i mëposhtëm i statistikës Λ realizon pikerisht , dhenien e pergjigjes se testit, duke përdorur shpërndarjen X^2. Transformimi qe ai propozoi është:

$$-(n - 1 - \frac{1}{2}(p + k))\ln \Lambda$$

dhe statistika e formuar ka shpërndarje X^2 me p(k-1)shkallë lirie.
Idea e perafrimit te shperndarjes në shqyrtim me shpërndarjen X^2 me p(k-1)shkallë lirie është bërë nga Barlett që më 1947. Ky modifikim I statistikës përdoret veçanerisht nëse nuk jemi të treguar të kujdesshem për testin e normalizimit të variablit të pavarur Y.

4.1.5.1 Vërejtje

Programet kompjuterike të cilat kryejnë metodën MANOVA, të programuar e modeluar nga Philip Burns nga Universiteti I NorthWest-it, bëjnë Analizen e Kovariancës Shumëpërmasore dhe Analizë të Variancës Njëpërmasore për çdo

variabël, si dhe analizën e regresit. Për çdo efekt të testuar afishohet nje faqe output-i. Pra afishohen tri output-e

- Testet e Variancës së Shumëfishtë. Këto teste përfshijnë kriterin llambda Λ Wilk-i që përafrohet duke shfrytëzuar përafrimin e Rao-s me shpërndarjen *F*, si dhe teste të tjerë të përmendura në paragrafin e mësipërm.
- Testet ANOVA një-përmasore për çdo variabël, të cilat bëjnë të mundur përcaktimin e ndikimit apo jo të variablit të pavarur në secilin prej variablave të varur në studim.
- t-testet të cilat kanë vlerë nëse shihet ndikim I variablit të pavarur tek variablat e varur, për të kuptuar ç'nivele të variablit të pavarur ndikojnë tek këto variabla.

4.1.5.2 Zbatimi i 4.1 (vazhdim)

Duke vrojtuar me kujdes tri output-et e MANOVA-s (shih problemin Ndikimi I nivelit te ndotjes ne problemet shendetësore të përgjithshme, ne problemet dermatologjike dhe ato pneumatike), arrihet në dheninien e disa konkluzioneve.

Nga Analiza e Variancës së Shumëfishtë, meqënëse jemi në rastin kur nuk i është kushtuar kujdes normalizimit të të dhënave, por ato janë konsideruar normale nisur nga vlerat e analizes deskriptive të variablave, lexohet vlera e statistikës Fisher nga afishimi I Pillai's Trace, dhe kjo është 4.654 dhe i korespondon p- vlera 1.2 E-12. Kjo tregon se ka efekt variabli I pavarur , ne rastin tonë *'fshatrat'*, në vlerat e gjithe variablave të varura, kur ato të 13 shihen si një grup I vetëm. Por ky ndikim mund të jetë në një nëngrup të këtyrë 13 variablave te varur.

Nga Testet ANOVA një-përmasore për çdo variabël, Tabela 4.1, vërehet ndikim I variablit *'fshatrat'* në secilin prej 13 variablat e varur, të cilat mund të shikohen të përcaktuara ne tabelën në fjalë.

Për të përcaktuar se midis cilave vlerave te variablit te pavarur '*fshatrat*' vërehet ndryshim, janë kryer t-testet për secilën nga dy çiftet e zonave në shqyrtim, për çdo variabel të varur.

Tabela 4–2 ANOVA për 13 variablat e varur

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
fshatrat	sy te perlotur	4.706	2	2.353	10.392	.000	.110
	kruarje hundesh	1.918	2	.959	5.610	.004	.063
	tharje gryke	10.481	2	5.240	27.313	.000	.245
	gelbaze	4.852	2	2.426	11.020	.000	.116
	kolle	2.674	2	1.337	5.935	.003	.066
	lunga ne lekure	.905	2	.452	4.755	.010	.054
	kruarje ne lekure	2.357	2	1.179	9.113	.000	.098
	shperthim puçrrash apo flluskash	1.404	2	.702	6.373	.002	.071
	skuqje nxehtesi ne fytyre	14.098	2	7.049	44.434	.000	.346

	pezmatime te lekures	3.556	2	1.778	12.302	.000	.128
	marramendje	6.952	2	3.476	16.507	.000	.164
	nauze, te vjella	5.380	2	2.690	12.306	.000	.128
	dhimbje koke te forta	13.601	2	6.800	45.622	.000	.352

. Duke u perqendruar në vlerat e t-testeve te afishuara mund të arrihet në perfundimin që

- Ka ndryshim signifikativ midis fshatrave Kurjan dhe Sheqishtë- Beline dhe Kurjan e Marinez, për çdo pyetje te bëre. Vlera e t-testit rezulton pozitiv, ndaj themi se vlerat e këtyre variablave janë më të larta në Kurjan. Duke vërejtur mënyrën e kodimit të përgjigjeve në fjalë, konkludohet se ka ndikim negativ të zonës industriale në tërë parametrat e përcaktuara si percaktues të gjendjes së përgjithshme, të gjendjes pneumatike dhe asaj dermatologjike.
- Nuk vihet re ndryshim midis dy fshatrave që ndodhen në zonat industriale. Mesataret e nënpopullimeve përkatëse konsiderohen të njejta.

4.2 Analiza e Komponentëve Kryesorë

Le të jenë $X_1,X_2,\ldots X_p$ p variabla të rastit që kanë shpërndarje shumë përmasore, jo dosmosdoshmërisht normale me pritje matematike $\mu^{p*1} = (\mu_1, \mu_{2,\ldots}\mu_{1p})'$ dhe matricë kovariance . Zakonishti tregohet shume kujdes në aplikime studimit të kesaj matrice kovariance, pra studimit të lidhjes mes variablave $X_1,X_2,\ldots X_p$. Studimi I kësaj lidhje ndryshe, njihet me emrin, studimi I *struktures së variancës*. Struktura e variancës mund të përcaktohet ose nga llogaritja e kovariancës mes variablave, ose e koorelacionit midis tyre.

Qëllimi i këtij aplikacioni është të gjenden, nëse është e mundur, disa funksione $Y_1,Y_2,\ldots Y_q$, ku çdo Y_i është kombinim linear i $X_1,X_2,\ldots X_p$ ku (q<p), i cili afersisht gjeneron strukturën mes $X_1,X_2,\ldots X_p$. Për me tepër të gjendet strukturë e re varësie, që kjo të mund të sigurojë të njëjtën sasi informacioni si variablat origjinalë.

Më poshtë paraqitet një metodë që njihet me emrin *analiza e komponentëve kryesore*, që tregon pikërisht parimet e punës së kësaj analize.

Metoda synon të gjejë p kombinime lineare

$Y_1 = \sum_{j=1}^{P} \alpha_{1j}X_j, \ldots \ldots, Y_p = \sum_{j=1}^{P} \alpha_{pj}X_j$

të tillë që

cov(Yi, Yj)=0 për i,j= 1,…, p nese i≠j,

$D(Y_1) \geq D(Y_2) \geq \ldots \geq D(Y_p)$

dhe

$$\sum_{j=1}^{P} D(Y_j) = \sum_{j=1}^{P} \sigma_{ii}$$

Nga barazimet e mësiperme duket sikur variablat e rinj jane të pakooreluar dhe të radhitur sipas dispersioneve, kovariancës së tyre. Për më tepër *totali I variancës mbetet I njëjtë* pas transformimit, në këtë mënyrë, nënbashkësia e q variablave të rinj të parë, Y_i per i=1,2,. . .,q mund të shpjegojë shumicën e variancës totale dhe për më tepër të sigurojë një përshkrim të strukturës së varësisë mes variablave origjinalë. Për përcaktimin e koeficientëve $\alpha_{ij,}$për i,j = 1, . . .p përdoret metoda e komponentëve kryesorë.

Shpërndarja e variablave $X_1,X_2,\ldots X_p$ përcaktohet tërësisht nga **μ** dhe **Σ.** Me supozimin që **μ**=(0,0,…0)' , struktura e varësisë mes variablave që jipet nga **Σ**, shpjegon tërësisht shpërndarjen e $X_1,X_2,\ldots X_p$.

Duke ditur që **Σ** është e njohur , le ta kerkojme $Y_1= \alpha_{11}X_1+. . . + \alpha_{1p}X_p$. Duhet të gjejmë $\alpha_{11,} . . \alpha_{1p,}$ të tillë që

$$D(Y_1) = \sum_{i=1}^{P}\sum_{j=1}^{p} \alpha_{1i}\,\alpha_{1j}\sigma_{ij}$$

te ketë vlerë maksimale dhe të plotësohet kushtit që vendoset për të siguruar unicitetin e zgjidhjes, $\sum_{j=1}^{P} \alpha_{1j}{}^2 = 1$

Zgjidhja $\alpha_1=(\alpha_{11}\ \alpha_{12},. . . \alpha_{1p})'$ quhen vektorë të vetë dhe shoqërohet me me të madhin vlerë te vetë të **Σ.** Kjo vlerë e vektorit të vetë është e barabartë me variancën D(Y1) kombinimi linear $Y_1= \alpha_{11}X_1+. . . + \alpha_{1p}X_{1p}$ quhet edhe komponenti I parë kryesor I komponentëve të $X_1,X_2,\ldots X_{p,}$ dhe ai shpjegon $100*D(Y_1)/V$ përqind të variancës totale.

Pastaj, kërkojme të përcaktojme $Y_2= \alpha_{21}X_1+. . . + \alpha_{2p}X_p$, dhe të vlerësojme $\alpha_{21,}\ \alpha_{22,}. . \alpha_{2p.}$

Përcaktimi I këtyre koeficientëve behet I tille qe të maksimizojë

$$D(Y_2) = \sum_{i=1}^{P}\sum_{j=1}^{p} \alpha_{2i}\,\alpha_{2j}\sigma_{ij}$$

dhe $\sum_{j=1}^{P}\left(\alpha_{2j}\right)^2 = 1$ (*) si dhe $\text{cov}(Y1,Y2)= \sum_{i=1}^{P}\sum_{j=1}^{p} \alpha_{1i}\,\alpha_{2j}\sigma_{ij}=0$ (**)

Kushti (*)si guron unicitetin e zgjidhjes ndërsa (**)siguron qe Y1 dhe Y2 të jenë të pakooreluar. Zgjidhja $\alpha_2=(\alpha_{21}\ \alpha_{22},. . . \alpha_{2p})'$ është vektor I vetë dhe shoqërohet me vleren e vetë të dytën për nga madhesia e të **Σ.** Kjo vlerë e vetë është e barabartë me variancën D(Y2). Kombinimi linear $Y_2= \alpha_{21}X_1+. . . + \alpha_{2p}X_{1p}$ është dhe komponenti I dytë kryesor i komponentëve të $X_1,X_2,\ldots X_{p,}$ dhe, të dy komponentët e parë kryesorë shpjegojnë $100*(D(Y_1) + D(Y_2))/V$ përqind të variancës totale.

Pasi $Y_1, Y_2,. . ., Y_{q-1}$, për q=2,3,…,p të jenë marre në këtë menyrë fitohet $Y_q= \alpha_{q1}X_1+. . . + \alpha_{qp}X_p$ I tillë që

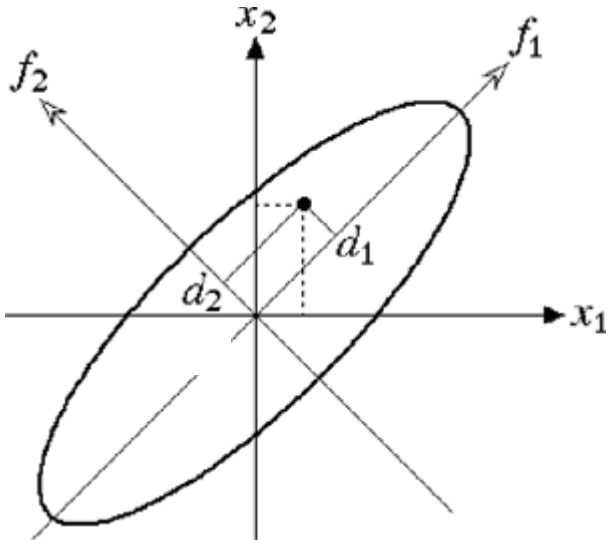
$$D(Y_q) = \sum_{i=1}^{P}\sum_{j=1}^{p} \alpha_{qi}\,\alpha_{qj}\sigma_{ij}$$

dhe që plotëson kushtin $\sum_{j=1}^{P} \alpha_{qj}{}^2 = 1$ dhe $\text{cov}(Y_m,Y_q)= \sum_{i=1}^{P}\sum_{j=1}^{p} \alpha_{qi}\,\alpha_{mj}\sigma_{ij}=0$ për m= 1,. . . q-1. Kështu fitojmë vektorin e vetë $\alpha_q=(\alpha_{q1}\ \alpha_{q2},. . . \alpha_{qp})'$ që I korespondon vlera e vetë e q-të më e madhe D(Yq). Kështu Yq është komponenti i q-të kryesor, dhe variablat $Y_1, Y_2, ... Y_q$ shpjegojnë $100*(D(Y_1) + D(Y_2) +. . . + D(Yq))/V$ përqind të variancës totale.

Një interpretim gjeometrik I problemit ne fjale mund të shihet në figuren, për p=2. Çdo variabel $X_1,X_2,\ldots X_{p,}$ paraqitet gjeometrikisht nga nje bosht kordinativ me origjine në $\mu^{p*1} = \left(\mu_1 , \mu_{2,\ldots}\mu_{1p}\right)'$. Këto p boshte në hapësirën p dimensionale, bejne të mundur konceptimin e

$X= (x_1, x_2, . . X_p)'$ si një pikë kordinatat e të cilës jane $X_1=x_1, X_2=x_2,\ldots X_p=x_p$.

Figura 4–1 Interpretimi gjeometrik i metodës së komponentëve kryesor në rastit n=2



Problemi i analizes së komponenteve, sipas kësaj paraqitje gjeometrike, mund të ishte i njëjtë si të rrotulluarit e këtyre boshteve që variabli Y_1 i paraqitur nga i pari bosht i ri të këtë dispersion maksimal. Variabli i dytë Y_2 i paraqitur ne boshtin tjetër të ri është i pakoreluar me Y1 dhe ka nje dispersion maksimal. Ne mënyrë të ngjashme , variabli Y_q paraqet boshtin e ri të q-të që është i pakooreluar me $Y_1, Y_2, \ldots, Y_{q-1}$ dhe ka maksimumin e dispersionit, pa e thyer kushtin e lartpërmendur. Kështu nëse f(x) do të ishte nje densitet normal i vektoreve te rastit $X= (x_1, x_2, \ldots X_p)'$, atehere inekuacioni $f(x) \leq c$, ku c është një konstante, përcakton si të thuash në hapësirën p-dimensionale nje elipsoid . Drejtimi I boshteve kryesore të ketij elipsoidi përcakton dhe drejtimin e komponenteve të rinj Y_i.

I parë ne hapësirën dy dimensionale të përcaktuar nga X_1 dhe X_2 me origjinë në (μ_1 , μ_2)

elipsoidi është elipsoid I zakonshëm. Komponentja e parë është në drejtimin e boshtit të madh të elipsit, ndersa komponentja e dytë kryesore është në drejtim të boshtit të vogël të elipsit.

Pra, problemi i përcaktimit të komponentëve kryesorë të rinj bazohet ne matricën e kovariances Σ. Nëse kjo matricë është e panjohur atëhere në vend të saj do të përdoret matrica e kovariancës së zgjedhjes.

4.2.1 Vërejtje

1. Kryerja e metodes MANOVA ne programet kompjuterike, si metode me shume variabla behet nëpermjet analizes 'multivariate'. Dhe kjo ka arsyen e saj, theksuam se ne kete metode kemi shume variabla te varur dhe nje variabel ne dukje I pavarur me to. Por, sipas Tabachnik dhe Fidell (2007)[14] është e pakuptimtë përdorimi I MANOVA-s nëse të gjithë variablat e varur masin të njëjtën gje. Ato duhet të mendohet se I përkasin, të paktën dy grupimeve të ndryshme.

2. Ne dallim me procedurat e tjera te analizës shumë përmasore, të cilat kanë si parim reduktimin e numrit të variablave të nevojshëm për tu matur, procedura në fjalë

kërkon që të gjithë p variablat të maten, sepse ato të gjithë do të ndikojnë në llogaritjen e vlerës së komponenteve kryesorë. Pra, është e pakuptimtë të perdoret nëse qëllimi I studimit është se cilët nga variablat e duhet të hiqen, që të marr më pak variabla në studim.

3. Homogjeniteti I variancës me grupeve është shumë I rëndësishëm. Siç dihet kontrolli në fjalë bëhet përmes testit Levene. Metoda interesohet që varianca mes grupeve të jetë e njejtë, por, duhet theksuar, se ky kusht do të konsiderohet që nuk plotësohet kur p-vlera e afishuar të jetë <0.005, jo si zakonisht < 0.05.[13]

4. Në qoftëse $X_1,X_2,\ldots X_p$ kanë shpërndarje normale shumëpermasore, atehere komponentët kryesorë $Y_1,Y_2,\ldots Y_q$, duke qenë se jane çift e çift te pakooreluar, do të jene çift e çift normale.

5. Vërehet se varianca totale e komponentëve kryesorë të rinj është sa varianca mes variablave të varur. Por, duke patur parasysh se për çdo variabël të varur X_i mund të behet normalizimi i tij , varianca totale V do të jetë sa numri I variablave të varur X_i .

6. Njehsimi I koorelacionit mes X_i dhe Yj jipet nga $\frac{\alpha_{ij}(D(Y_j))^{1/2}}{\sigma(Xi)}$, ndaj për të vlerësuar kontributin e $X_1,X_2,\ldots X_p$ në Y_j, përdoren pikërisht $\frac{\alpha_{ij}}{\sigma(Xi)}$. Po nëse vlerësimi behet duke u bazuar tek kovariancat atëhere përdoren vetëm α_{ij} për vlerësimin e komponentëve Xi që ndikojnë në përcaktimin e komponentit të ri kryesor. Kur ky koeficientë është më i lartë, atëhere më i lartë është kontributi I variablit te varur.

4.2.2 Zbatimi 2(*vazhdim*)

Po duhet theksuar se vlera me e madhe e metodës MANOVA verehet në mënyrën tjetër që mund të kryhet kjo metode, në analizën e komponenteve kryesore. Përcaktimi I ketyre komponentëve kryesor është një përcaktim i i variablave të rinj që janë kombinim linear i 13 variablave te varur ne studim,e shpesh për këtë arsye njihen me emrin përcaktimi i supervariablave. Këto variabla të rinj, supervariabla[15], të pavarur synojnë të paraqesin të njëjtat matje apo informacione të marra nga 13 variablat në studim.

MANOVA për analizën e komponentëve kryesorë do te ekzekutohet pas shkrimit të kodit në editorin e sintaksës. Editori është emertuar ' *pyetje'* dhe permban kodin e meposhtëm:

```
MANOVA p23 p24 p25 gelbaze kolle p28 p29 p30 p31 p32 p33 p34 p35 BY
fshatrat(1,3)
/DISCRIM=STAN RAW CORR
/PRINT =SIGNIF( MULTIV,UNIV,EIGEN,DIMENR)/DESIGN.
```

Ekzekutimi i kësaj analize siguron gjetjen e dy kombinimeve lineare të 13 variablave në studim. Shihet se keto dy supervariabla të krijuar, ruajnë totalin e variances.

Tabela 4–3 Vlerat e veta të komponentëve kryesorë

Root No	Eigenvalue.	Pct.	Cum. Pct.	Canon Cor.
1	1.2537	100.00000	100.00000	.74585
2	.00000	.00000	100.00000	.00000

Por, afishimet e Tabeles 4.3, tregojnë se supervariabli i pare shpjegon 0.74585^2= 0.556% të variancës totale . Ndërsa supervariabli I dytë i korespondon nje vlerë e vete shumë e vogël 2.3E-5 dhe, për rrjedhojë dhe sasia e variancës që shpjegon supervariabli I dytë, siç shihet edhe nga afishimet e tabelës 1, është e papërfillshme (<.005). Kështu studimi përqëndrohet vetëm ne supervariablin e parë me vlerë të vetë 1.25375.

superndikimi= -.31412* p23 -.23080*p24-.50926*p25-.32348* gelbaze -3739* kolle -.21248*p28 -.29417*p29-.24600* p30-.64955* p31-.34177* p32-.39591* p33-.34184* p34-.65818* p35

Në rastin në fjalë problemi i studimit të ndikimit të 'fshatra'-ve tek 13 variablat e varur reduktohet ne studimin e 'fshatra'-ve tek variabli kanonik '*superndikim*'.

4.3 Analiza Faktoriale

Në seksioni 4.2 u tregua se teknika e komponenteve kryesore ishte një metode që bazohej në varësinë mes $X_1,X_2,\ldots X_p$ p variabla të rastit që kanë shpërndarje shumë përmasore, me pritje matematike $\mu^{p*1} = (\mu_1, \mu_{2,\ldots}\mu_{1p})'$ dhe matricë kovariance Σ^{p*p}. Komponentët kryesorë u kërkuan si p kombinime lineare

$Y_1 = \sum_{j=1}^{P} \alpha_{1j} X_j, \ldots, Y_p = \sum_{j=1}^{P} \alpha_{pj} X_j$ (*)

të variablave të pakooreluar $X_1,X_2,\ldots X_p$,dhe rradhiten sipas dispersionit të tyre

$D(Y_1) \geq D(Y_2) \geq \ldots \geq D(Y_p)$

dhe

$$\sum_{j=1}^{P} D(Y_j) = \sum_{j=1}^{P} \sigma_{ii} = V$$

Por, për ekuacionet (*) si ekuacione lineare, mund të shtrohet dhe problemi I anasjelltë, pra që variablat e varur të shprehen si kombinim linear I Komponentëve kryesorë. Pra

$X_1 = \sum_{j=1}^{P} \beta_{1j} Y_j, \ldots, X_p = \sum_{j=1}^{P} \beta_{pj} Y_j$

për disa konstante β_{ij}, i,j=1,2,..,p. Mund të tregohet[5] që $\beta_{ij} = \alpha_{ji}$, ndaj $X_1 = \sum_{j=1}^{P} \alpha_{j1} Y_j$, $\ldots, X_p = \sum_{j=1}^{P} \alpha_{jp} Y_j$

Nga këto ekuacione Modeli I Komponentëve Kryesorë qe plotëson kushtin

$$D(Y_2) = \sum_{i=1}^{P} \sum_{j=1}^{p} \alpha_{2i}\, \alpha_{2j} \sigma_{ij} = \sum_{i=1}^{P} \sum_{j=1}^{p} \alpha_{2i}\, \alpha_{2j} \sigma_{ij}$$

tani merr formën $\sigma_{ij} = \sum_{k=1}^{P} \alpha_{ki}\, D(Y_k) \alpha_{kj}$

$\sigma_{ii} = \sum_{k=1}^{P} \alpha_{ki}^2 \quad D(Y_k) pwr\ i.j = 1,2,\ldots p$

Këto ekuacione paraqesin faktorizimin e variancës dhe kovariancës së variablave të varur si funksion te α_{ij} dhe të Variances së Variablave Kryesorë.

Analiza faktoriale është një disiplinë e tërë, por qëllimi I këtij seksioni nuk është të trajtojë problemin në tërësi. Qëllimi është të shohim se si kjo metode na ndihmon në përcaktimin e Faktorëve Kryesorë. Programet statistikore kompjuterike kane paketa të gatshme për kryerjen e kësaj metode, ndonëse ne dukje ajo është shumë e vështirë.

Vlerësimi I parë I kësaj metode është bërë që më 1967 nga Morrisin[16], I cili sygjeroi që për përcaktimin e Faktorëve Kryesorë, të punohej me teknikën e rrotullimit. Paketat e programeve kompjuterike specifikojne:

- numrin e Faktorëve Kryesorë që mjafton të studiohen
- formen e matricës , që përcakton këto faktorë kryesorë
- vlerëson numrin e hapave të nevojshëm që ky vlerësim të bëhet I vlefshëm.

Nëse shtrohet problemii numrit të faktorëve kryesorë që është i nevojshëm të merren në shqyrtim sygjerohet[17]

- të merren faktorë që vlerat e veta t'i kenë >1
- këto komponentë në tërësi justifikojne së bashku rreth 70-75% të variancës totale.

4.3.1 Zbatimi 3(*vazhdim*)

Nëse shembullin e ndikimit te '*fshatra*'-ve tek 13 pyetjet qe percaktonin 13 variabla kategorikë, e shohim sipas Analizës së Komponentëve Kryesore, duhet të përmëndim që 13 pyetjet ndahen në tri kategori, në pyetje për gjendjen e përgjithshme të individit, për gjendjen pneumatike dhe dermatologjike. Kryerja e kësaj analize afishon tabelen e vlerave të veta, Tabela 4.4, që në bazë të verejtjes së mësipërme tregon se numri I faktorëve kryesorë duhet të jetë 3, e këto tri faktorë së bashku justifikojnë 58.9% të variancës totale.

Tabela 4–4 Tabela e vlerave të veta

Total Variance Explained

Comp.	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.313	33.174	33.174	4.313	33.174	33.174	3.091	23.776	23.776
2	2.103	16.179	49.353	2.103	16.179	49.353	2.614	20.111	43.887
3	1.231	9.466	58.820	1.231	9.466	58.820	1.941	14.933	58.820
4	.988	7.602	66.421						
5	.885	6.804	73.225						
6	.793	6.097	79.322						
7	.618	4.754	84.076						
8	.493	3.790	87.866						
9	.450	3.463	91.329						
10	.368	2.828	94.157						
11	.311	2.390	96.547						
12	.264	2.035	98.582						
13	.184	1.418	100.000						

Duke u mbeshtetur ne vlerat e afishuara nga kjo metodë për përcaktimin e faktorëve arrihen të ndertohen tri faktorët ne editorin e sintakses, si mëposhtë.

Komp_dermatologjik= .324* p23+ .460*p24+.312*p25+.142* gelbaze-.062* kolle +.834*p28 +.773*p29+.872* p30+.020* p31+.692* p32-.168* p33+.161* p34+.256*p35

Komp_pergjith= *.542* p23+ .273*p24+.614*p25+.098* gelbaze+.063* kolle +.026*p28 +.165*p29+.104* p30+.437* p31+.108* p32+.830* p33+.576* p34+.770*p35*

Komp_pneumologjik= *122* p23+.178*p24+.405*p25+.730* gelbaze+.783* kolle -.084*p28 +.034*p29-.068* p30+.600* p31+.310* p32-.104* p33+.275* p34+.167*p35*

Theksojme se faktoret e ndertuar janë kombinim linear i 13 variablave të varur, por nisur nga vlera me e lartë e koefiçientëve prane pyetjeve të karakterit dermatologjik në faktorin e parë, të pyetjeve të karakterit të përgjithshëm në faktorin e dytë dhe atij pneumologjik në të tretin është bërë dhe emërtimi I faktorëve. Të tri faktoret jane përberës të konsiderueshëm te variacionit të përgjithshëm, ndaj emertohen *komponente. (shih problemin 1, Kapitulli 6)*

Pastaj studiohet problemi I ndikimit të *'fshata'*-ve në këto tri faktorë kryesore.

5 Modeli i regresit logjistik të shumëfishtë

5.1 Regresi logjistik dhe llojet e tij

Në seksionin 2 është trajtuar metoda e regresit të shumefishtë linear (GLM). Por, lidhjet mund të jenë më të komplikuara se

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \varepsilon_i \qquad \text{për } i = 1..n$$

ku duhet të vlerësonim β_0, β_1, β_2 ,...β_p që janë parametrat e këtij modeli dhe ϵ_1, $\epsilon_{,2}$... $\epsilon_{n,}$, n variabla të pavarur $N(0,\sigma^2)$. Në këtë seksion do të përqëndrohemi në modelin e regresionit logjistik, duke trajtuar problemin e e llojshmërisë së tij në përputhje me problemin e shtruar, vlerësimet që ato sigurojnë dhe, më pas, mënyren se si ky model kontrollohet per vlefshmerine e tij.

5.1.1 Modeli i regresit logjistik të shumëfishte

Një model regresi përmban tri komponentë: komponentin e rastit, komponentin sistematik dhe funksionin lidhës.
Komponentët e rastit jane Y_1, Y_2,…Y_n që mendohet se janë ndryshore rasti të pavarura, me shpërndarje nga familja eksponenciale. Pra, shpërndarja e Y_i nuk është identike, por i takon të njëjtës familje eksponenciale.
Komponenti sistematik është modeli. Ai është funksion i parametrave.

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}$$

Në modelin e regresit linear, është i lidhur me mesataren e Y_i.
Dhe së fundmi, *funksioni lidhës* i dy komponentëve, që teorikisht paraqitet:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}, \text{ ku } \mu_i = E(Y_i).$$

Dimë se një ndër kushtet që duhet të plotësoheshin në këtë model ishte,Y_i të kishin shpërndarje normale. Por, meqënëse familja normale është pjesë e asaj eksponenciale atëhere ky lloj modelimi u përdor edhe kur Y_i kishin shpërndarje nga e njëjta familje , jo vetëm normale, por dhe Puasoni e Binomiale. Në rastet kur komponenti i rastit merr një numër të kufizuar vlerash, propozohet të përdoret metoda e modelimit të regresit logjistik.
Në këtë model , kur variablat e rastit Y_1, Y_2,… Y_n, janë të pavarur dhe kanë shpërndarje Bernuli (pi), lidhja mes x_i dhe p_i kërkohet në formen

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} \qquad = \text{logit}(p_i) \qquad (1)$$

Ana e majte paraqet *ln* e mundesive të suksesit për Y_i (që njihet me emrin *logit*)[18]. Modeli, siç shihet, propozon qe *ln* e mundësive të suksesit të jetë funksion linear i x-it.
Densiteti probabilitar i vektorit të rastit, në këtë rast, mund të shkruhet në formën:

$$p^y(1-p)^{1-y} = (1-p)exp\{y\ ln(\ p_i/(1-p_{i\_})\ \ \}$$

Termi $\ln(\frac{p}{1-p})$është parameter natyror i familjes eksponenciale, dhe siç shihet si funksion lidhës është përdorur g(p)= $\ln(\frac{p}{1-p})$.
Ekuacioni (1)mund të shkruhet si:

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}}{1+e^{\beta_0+\beta_1 x i_1+\cdots+\beta_n x_{in}}} \qquad (2)$$

ose. në mënyrë më të përgjithshme, si:

$$p(x) \ = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} \qquad (3)$$

Shihet se 0<p(x)<1 e në rastin në fjale, kjo përputhet me faktin që, ai është probabiliteti.

5.1.2 Modeli i regresit logjistik i thjeshtë me një variabël në modelim

Si lloj i vecante i regresit logjistik mund te konsiderohet edhe regresi logjistik I thjeshtë, që eshtë rasti I regresit logjistik të shumëfishtë , por me një variabël në modelim. Kjo është forma më e thjeshte e regresit logjistik[19].
Ai merr rendesi edhe teorike, për të kuptuarit më mire të komponenteve të këtij regresi dhe interpretimin korrekt të tyre. Pra edhe për lehtësi në studim e p(x), komponentin sistematik mund ta paraqesim, duke e kufizuar në 1 numrin e variablave që marrin pjesë në modelim.
Atehere (3) do merrte formën

$$p(x) \ = \frac{e^{\beta_0+\beta\ \mathrm{x}}}{1+e^{\beta_0+\beta\ \mathrm{x}}}$$

dhe derivati mund të shkruhej:

$$\frac{dp(x)}{dx} = \beta p(x)\big(1-p(x)\big) \qquad (4)$$

Meqë termi $p(x)\big(1-p(x)\big) > 0$, atehere shenja e derivatit varet nga shenja e β. Nqs β>0 atëhere derivati i p(x) është rigorozisht pozitiv; nqs β<0 atëhere derivati p(x) është rigorozisht negativ. Pra, nqs β>0 atëhere p(x) është funksion rigorozisht rritës; nqs β<0 atëhere p(x) është funksion rigorozisht zbrites; nqs β=0 atëhere p(x) =$\frac{e^{\beta_0}}{1+e^{\beta_0}}$ për çdo x. Në këtë rast si në modelin e regresit linear, nuk ka lidhje mes *p* dhe *x*.
Parametrat β_0 dhe β kanë pothuajse kuptim të ngjashëm me ato te regresit linear të zakonshëm.
Nëse x=0 , po të shihet ekuacioni(1) kuptohet se, β_0 paraqet *ln* e raportit të mundësisë së suksesit ne x=0. Nëse duke zëvendesuar tek (1) me x dhe x+1 kemi se për çdo x është e vërtetë:

$$\ln\frac{p(x+1)}{1-p(x+1)} - ln\frac{p(x)}{1-p(x)} = \beta_0 + \beta(x+1) - \beta_0 - \beta(x) = \beta$$

Kështu β ndryshimi i *ln* së mundësive të suksesit që ndodh nëse x rritet me 1 njësi.(*)
Duke marrë eksponencialin e të dy anëve të këtij ekuacioni, fitohet

$$e^{\beta} = \frac{p(x+1)/(1-p(x+1))}{p(x)/(1-p(x))} \quad (5)$$

Ana e djathte paraqet raportin e mundesive dhe krahason mundësitë e suksesit ne x +1 me ato te suksesit ne x [21]. Përfundimisht,

$$\frac{p(x+1)}{1-p(x+1)} = e^{\beta}\frac{p(x)}{1-p(x)}. \quad (6)$$

Pra, e^{β} është faktori që lidh ndryshimet e mundesisë së suksesit, që i korespondojnë rritjes me një njësi të x-it.
Ekuacioni (3) sygjeron një tjetër mundësi te modelimit të probabilitetit të Bernulit p(x) si funksion i x-it. Dihet se densiteti probabilitar i n.r. *logistic(0,1)* përcaktohet nga barazimi

$$F(w)=\frac{e^w}{(1+e^w)}$$

Kështu, barazimi (3) mund të shkruhet në formën:

$$p(x) = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p)$$

Interpretimi i dhënë tek (*) është një interpretim mese i saktë, por nëse e shohim në këndvështrimin që p(x) paraqet një kurbë, duke u bazuar në ekuacionin (4) ,themi se nëse ndërtohet tangentja me kurbaturën[20], ne një pikë x , koeficienti këndor i tangentes do të jetë $\beta * p(x)(1-p(x))$.
Kështu për p(x)=1/2 , ky koeficient e ka vlerën $\beta * \frac{1}{2} * \frac{1}{2} = \beta/4$.
Shihet se p(x)=1/2, arrihet në x= - β_0/β.
Kjo vlerë e x konsiderohet edhe si niveli i efektit të medianës. Në studimet mjekësore kjo quhet ndryshe doza e rrezikshme shkatërruese [23].
Duke u nisur nga perafrimi linear i kurbes, Agresti[22] sygjeron të vrojtohet se ç'ndodh kur kemi një ndryshim të x me $\frac{1}{\beta}$.
Këtij ndyshimi i korespondon një ndryshim afërsisht sa $\frac{1}{\beta} * \frac{\beta}{4}$=1/4 i p(x). Pra pikat ku p(x)=0.25 dhe 0.75 (në të vërtetë kjo ndodh ne 0.27 dhe 0.73) dhe në 0.50, jane me distancë $\frac{1}{\beta}$ nga njëra tjetra. (**)
Kur p(x)=0.9 ose 0.1, koeficienti kendor do të jete 0.09 β.

5.1.2.1 Përfundime

Interpretimi i koeficientëve në rastin logjistik të thjeshtë me një variabël në modelim, është:

- kur x rritet me nje njësi, β paraqet ndryshimin e *ln* së mundësive të suksesit, ndërsa

- e^{β} është faktori që lidh ndryshimet e mundesisë së suksesit.

5.1.3 Modeli logjistik i thjeshtë me parametra kategorikë

Regresi logjistik, e ndërton modelin edhe nëse variabli që duhet të modelojmë është kategorik.
Për të thjeshtuar paraqitjen e problemit , supozohet se duam të punojmë me një parametër të tillë të vetëm , me *I* kategori, pra në rreshtin e *i*-të të tabelës përmbledhese të të dhënave do të ishte y_i, numri i ndodhjes së vlerës përkatëse të variablit që duam të modelojmë, gjatë n_i rasteve që shfaqet kjo vlerë e variablit kategorik.Y_i, mund të shihet si variabël binomial me parametër p_i. Modeli logjistik i tij me një faktor është:

$$\ln\frac{p_i}{1-p_i} = \beta_0 + \beta_i \quad (7)$$

Sa më e lartë β_i aq më e lartë vlera e p_i,

Po ana e djathtë e modelit të mësipërm është modeli ANOVA njëpërmasor. Si në metodën në fjalë modeli ka aq parametra sa nivele ka faktori i shqyrtuar. Por, këtu një nga nivelet e faktorit është reduktuar. Pra faktori i shqyrtuar me *I* kategori ka tani *i-1* kategori jo të reduktuara dhe kategorine e *I* te reduktuar. Një ndër parametrat merret kështu 0. Nëse $\beta_I = 0$ atëhere β_0 është logit në rreshtin I (zëvendesoj tek (7) indeksin $i = I$ dhe fitoj: $\text{logit}(p_I) = \beta_0$)

Ndaj $\beta_i$është diferenca mes *logit* të rreshtit të i dhe I. Kështu $\beta_i$është *ln* e raportit të mundësive , për këtë çift radhësh.
Nëse për faktorin në fjale ekziston nje *I* që p_i të jetë pozitive, pra të ekzistojë $\beta_i \neq 0$, atëhere modeli është i kënaqshëm[25]. Ndërsa nëse të gjitha $\beta_i = 0$, atëhere themi se X lidhet me Y vetëm një term konstant.

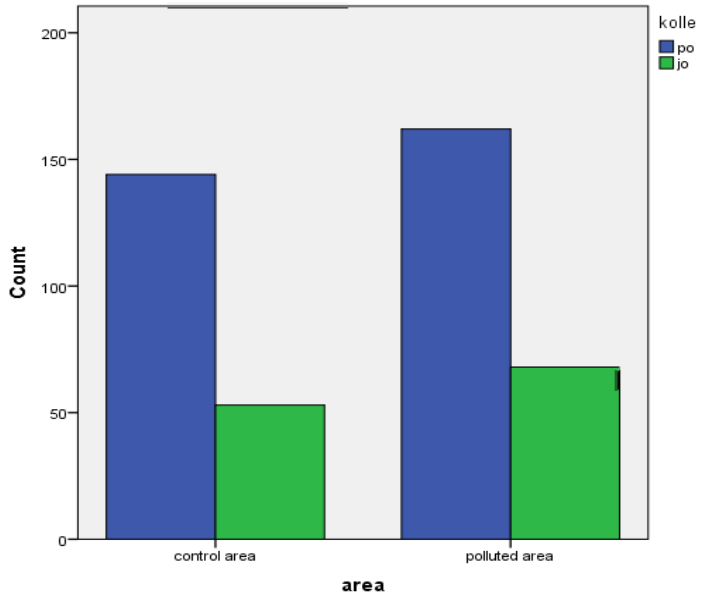
5.1.3.1 Përfundime

- Regresioni logjistik përdoret në situata kur duam të parashikojmë prezencën ose mungesën e një karakteristike në subjektet, individët, duke u bazuar në vlerat e një bashkësie variablash të studiuara mbi to .
- Kjo metodë pranon, që variablat në bazë të së cilës ndërtohet modeli, të jenë variabla kategorikë dhe të vazhdueshëm.
- Nëse tërë variablat që marrin pjesë ne modelim janë kategorikë, metoda është shumë efektive.
- Nëse ndonjë nga variablat e modelit është i vazhdueshem, ai mund të përcaktohet 'covariate', dhe modelimi bëhet duke punuar me variablin si në rastin e Analizës së Kovariancës (seksioni 3.3)
- Koeficientët e kësaj metode përdoren për të vlerësuar raportet e mundësive për secilin nga variablat e pavarur në model, nëse individi, subjekti i takon një grupi të përcaktuar nga vlera të caktuara të variablave të tjerë, që marrin pjesë në modelim.

5.1.3.2 Zbatim 1

Në problemin 3 Kapitulli 6, **Trajtimi i problemeve pneumologjike,** nga analiza e treguesve të përzgjedhur ne këtë fushë u vu re se treguesi kryesor është treguesi 'kollë'.
Kërkohet, në fakt, të kontrollohet nëse ndikon apo jo *zona* në mënyrë të dukshme në këtë parametër pneumatik.
Për të analizuar problemin në fjalë janë shfrytëzuar të dhënat e 426 subjekteve të marra rastësisht nga tri fshatrat e sipërpërmendura, 196 subjekte janë nga zona e pastër, zona e kontrollit, pjesa tjetër 230 janë marrë nga zona industriale, e konsideruar e ndotur.
Në bazën tonë të të dhënave, variabli *koll*ë nga transformimi i rekordeve, (shih Problemin 3), eshte i koduar, 0 nëse personi nuk ka kollë dhe 1 kur ka kollë. Pra është variabël binar.
Grafiket e paraqitur, Figura 5. 1, tregojnë për të dhenat e shfaqjes apo jo të ' kollës' ne dy zonat. Grafiket kane vetem karakter informues, ne kuptimin që, shihet se numri i rasteve me kollë është me i lartë në zonën e ndotur, por aty është më i lartë dhe numri i rasteve pakollë (vellimi i zgjedhjes ne zonen e ndotur eshte me i larte). Nga llogaritjet mundësitë e kollës në zonën e pastër janë 36.7%, ndërsa në zonën e ndotur jane 42%. Ndaj, shtrohet problemi, a është kjo gjë statistikisht e vërtetë.
Për të kontrolluar këtë, për shkak të natyrës së variablit 'kollë', është përdorur metoda e regresit logjistik binar me variabla kategorikë.
Variabel i varur është konsideruar variabli *'kollë'*, variabël binar; ndërsa variablat me të cilën është ndërtuar parashikimi është variabli *'zone' (area), 'konsum_duhan_në_shtëpi'* dhe variabli *'konsum_duhan'* , i trajtuar më parë si ndikues në *'kollë'* [26].

Figura 5–1 Treguesi 'kollë' sipas zonave



Vellimi i te zgjedhjes për shkak të mungesës së vlerave është 219, nga të cilët 112 nuk konsumojnë duhan pjesa tjetër konsumon. Kështu, nëse për nje person të zonës do të

na duhej të konkludonim konsumon apo jo duhan, pa bërë asnjë modilim do të thonim jo. Sipas tabelës së klasifikimit ky pohim do të ishte i vërtetë në 51.1% të rasteve.

Tabela 5–1 Vlerësimi para modelimit

Classification Table[a,b]

Observed		Predicted		
		kolle		
		jo	po	Percentage Correct
kolle	jo	112	0	100.0
	po	107	0	.0
Overall Percentage				51.1

a. Constant is included in the model.
b. The cut value is .500

Vlera exp(b)=0.955 tregon raportin individëve e me kolle ndaj pakolle, ndryshe 1-0.955=0.053 mund të interpretohetse 5% jane me shume rastet pa kolle se me kolle. Të gjitha këto vlerësime bëhen para modelimit. (në bllokun 0 të output-eve)

5.1.4 Regresioni logjistik multinominal

Regresi multinominal logjistik përdoret në shumë situata, kur duhet të klasifikohen subjektet bazuar në vlerat e një bashkësie variablash. Ky tip regresi është i ngjashem me regresin logjistik, por është më i përgjithshëm, sepse nuk kërkon që patjetër variabli i varur të ketë vetëm dy kategori.

5.1.4.1 Përfundimisht:

- Regresi multinominal logjistik përdoret kur duhet të klasifikohen subjektet bazuar në vlerat e një bashkësie variablash.
- Variablat e pavarur që marrin pjesë në model mund të jenë kategorik ose të vazhdueshëm.
- Nëse ato janë kategorik, konsiderohen si faktorë gjatë përpunimit të të dhënave, ndërsa nëse janë të vazhdueshëm, si 'covariates'
- Ky model ndërtohet, nëse pranohet se mundesitë e raportit të dy kategorive të variablit, që do të modelohet, nuk varen nga kategorite e tjera të variablit.

5.2 Vlerësimi i parametrave

Në modelin e regresit linear, ndërtohej modeli

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \varepsilon_i$$

e për vlerësimin e parametrave përdoret metoda e katrorëve më të vegjël. Ndërsa në rastin në fjalë kur Yi∼Bernuli(p_i), ato nuk kanë më lidhje direkte me

$\beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}$, kështu nuk mund të përdoret më metoda e katrorëve më të vegjël.

Metoda që përdoret për **vleresimin pikësor** është Metoda e Përngjasisë Maksimale. Ndërtohet funksioni I Përngjasisë Maksimale[25]

$$L(\beta_0, \dots, \beta_p) = \prod_{j=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} = \prod_{i=1}^{n} F_i^{y_i}\,(1 - F_i)^{1-y_i}:$$

Për te kontolluar ekstremumin e tij, maksimumin, përdorim funksionin ndihmës të Funksionit të Përngjasisë

$\ln(L(\beta_0, \dots, \beta_p)) = \sum_{i=1}^{n}\{\ln\,(1 - F_i) + y_i \ln\,(\frac{F_i}{1-F_i}) \quad \}$

dhe gjej derivatet e pjesshme.

Derivimi sipas β_0 siguron:

$$\frac{\partial \ln(1 - \mathrm{F_i})}{\partial \beta_0} = -\frac{f_i}{1 - F_i} = -\frac{F_i f_i}{F_i(1 - F_i)}$$

$$\frac{\partial \ln(\frac{\mathrm{F_i}}{1 - \mathrm{F_i}})}{\partial \beta_0} = \frac{f_i}{F_i(1 - F_i)}$$

Pra,

$$\frac{\partial \ln\,(L(\beta_0,\dots\beta_\mathrm{p}))}{\partial \beta_\mathrm{i}} = \sum_{i=1}^{n}(y_i - F_i)\frac{f_i}{F_i(1-F_i)} \quad (5)$$

Ndërsa derivimi sipas variablave të tjerë do të japë:

$$\frac{\partial \ln\,(L(\beta_0,\dots\beta_\mathrm{p}))}{\partial \beta_\mathrm{i}} = \sum_{i=1}^{n}(y_i - F_i)\frac{f_i}{F_i(1-F_i)}x_i \qquad (6)$$

Por, për funksionin logjistik, meqë $F(w) = \frac{e^w}{(1+e^w)}$

atëhere $\frac{f_i}{F_i(1-F_i)} = 1$

Ndaj, (5) dhe (6) marrin formë më të thjeshtë.

Vlerësimi I parametrave të modelit mund të behet duke barazuar me 0 ekuacionet (5) dhe (6) dhe, duke vlerësuar me anë të këtyre ekuacioneve β_0 dhe β_i .Ekuacionet e formuara janë jo linear dhe duhen zgjidhur numerikisht.

Për regresionin logjistik, funksioni ndihmes i Përngjasisë Maksimale është rigorozisht i myset, kështu nëse ekuacionet ne fjalë kanë zgjidhje. Kjo zgjidhje është e vetme, dhe është një vlerësim më Metodën e Perngjasise Maksimale. Megjithatë qëllon, që në raste të vecanta te dhënash, ekuacionet nuk kanë zgjidhje. Maksimumi arrihet në këto raste kur parametri x shkon në $\mp\infty$. Kjo lidhet me faktin se, gjatë modelimit logjistik, u supozua se $0<p(x)<1$, por në një bashkësi të dhënash të caktuara, maksimumi I Funksionit të Përngjasisë arrihet në limit, me $p(x)=0$ ose 1. Në këto raste thuhet se probabiliteti i të dhënave konvergjon në 0, nëse modeli logjistik është I vërtetë.

Ashtu si për vlerësimin e parametrave përdoret Metoda e Përngjasisë Maksimale, edhe për të vlerësuar **variancën** përdoret, si në rastin një përmasor përdoret

Informacioni I Fisherit, por në formë më të përgjithsuar, në atë të matricës së informacionit. Kjo matricë në rastin e dy parametrave ka formën:

$$\begin{pmatrix} -\frac{\partial^2 y}{\partial {\beta_0}^2}\ln L(\beta_0,\beta_1)/y) & -\frac{\partial^2 y}{\partial\beta_0\partial\beta_1}\ln L(\beta_0,\beta_1)/y) \\ -\frac{\partial^2 y}{\partial\beta_0\partial\beta_1}\ln L(\beta_0,\beta_1)/y) & -\frac{\partial^2 y}{\partial {\beta_1}^2}\ln L(\beta_0,\beta_1)/y) \end{pmatrix} \quad (7)$$

Duke punuar me ekuacionet (5) dhe (6) e regresit logjistik , matrica e informacionit (7) merr formen:

$$\begin{pmatrix} \sum_j n_j F_j(1-F_j) & \sum_j x_j n_j F_j(1-F_j) \\ \sum_j x_j n_j F_j(1-F_j) & \sum_j {x_j}^2 n_j F_j(1-F_j) \end{pmatrix} \quad (8)$$

dhe varianca e parametrave β_0, β_1 vlerësohen duke përdorur këtë matricë.[21]

Për llogaritjen e variancës ne baze të numrit të informacionit[22] $I(\tilde{\beta})$ perdoret formula,

Var $(\widetilde{\beta})) \approx [\ 1/I(\tilde{\beta})]$.Këtu në rastin shumë përmasor mund të përdorim, ne vend të inversit të informacionit të Fisherit, inversin e matricës së informacionit. Duke patur parasysh formen e matrisës inverse dypërmasore, kjo matricë e anasjelltë do të ketë në diagonalen kryesore elemente e diagonales kryesore të (7). Kështu, në këtë diagonale do të jenë vlerësuesit e $[\ se(\tilde{\beta}_0)]^2$dhe $[\ se(\tilde{\beta}_1)]^2$ ku $se(\tilde{\beta})$është devijimi standart i $\tilde{\beta}$.

Ashtu si në modelin e regresit linear , kontrollohen hipotezat Ho: β=0, pra a ka lidhje variabli që kërkojmë të modelojmë me variablat që marrin pjesë në modelim.

Për të kontrolluar këto hipoteza përdoret testi Wald. Ky test ka marrë emrin e statisticienit hungarez Abraham Wald. Sipas këtij testi parametrat vlerësohen nga vlerësimet e zgjedhjes.

Testi që përdoret për kontrollin e hipotezës në fjalë është :

$$\frac{\tilde{\beta}}{se(\beta)} \quad (9)$$

Ky test nëse është e vërtetë Ho ka shpërndarje afërsisht normale, vecanërisht kur vëllimi i zgjedhjes është i madh.

Si zakonisht, nëse vlera e vrojtuar e testit bie në zonën kritike , atëhere hipoteza bazë bie poshtë.

Ndryshe për vlerësimin e ndikimit apo jo të parametrit në modelim mund të përdoret testi[11]

$$-2\ln\lambda(y^*) = 2[lnL(\tilde{\beta}_0;\widetilde{\beta}_1|y^*) - lnL(\tilde{\beta}_0,0|y^*)] \quad (10)$$

ku $\tilde{\beta}_0$ është vlerësimi për β_0 kur $\beta = 0$

Tregohet se testi në fjalë ka shpërndarje X_1^2.

Kështu nëse vlera e vrojtuar e testit në fjalë plotëson kushtin -2lnλ ≥ $X^2_{1,nkr}$, atehere variabli në fjalë nuk ndikon në modelim.

5.2.1 Problemi i vlerësimit intervalor të parametrave:

Për modelin me vetëm një variabel sipas modelimit,
$logit[p(x)] = \alpha + \beta x$,
shqyrtuam kontrollin e hipotezës $H_0 : \beta = 0$. dhe testet me të cilat ky kontroll bëhet.
Kontrolli i saj behet, sic u përmend më sipër, me anë të **Testit Wald**-it që përdor funksionin logaritmik te përngjasisë maksimale për $\hat{\beta}$, me test: $Z = \hat{\beta}/SE$ ose **katrorin** e tij; kur H_0është e vërtete z^2 asimptotikisht ka shpërndarje χ^2_1.
Metoda **e maksimizimit të raportit të funksionit të përngjasisë maksimale për dy vlera**, e kthen problemin, duke punuar me funksionin logaritmik, në maksimizimin e diferencës ndërmjet logaritmit te funksionit të përngjasisë maksimale në $\hat{\beta}$ dhe në $\beta = 0$. E ky test (-2LL), gjithashtu ka asimptotikisht një shpërndarje χ^2_1.
Për zgjedhje me vëllim të madh, të tre testet zakonisht japin rezultate të ngjashme. Raporti Përngjasia Maksimale (-2LL) preferohet më shumë në krahasim me testin Wald. Merren me shumë informacione nëse përdoret kjo metode. Kur $|\beta|$është relativisht e madhe, testi Wald nuk është aq i fuqishëm sa test (-2LL), dhe mund edhe të cojë në përfundime jo normale [Hauck dhe Donner (1977)].

Intervalet e besueshmërisë japin më shumë informacion se sa kontrolli i hipotezave. Një interval për β sigurohet duke u pershtatur statistiken që do të duhej për vlerësimin e testit $H_0: \beta = \beta_0$. Intervali besimit është bashkësia e β_0 për të cilën testit me shpërndarje hi-katror merr vlerë jo më e madhe se $\chi^2_1(\alpha) = z^2_{\alpha/2}$. Sipas metodës Wald, kjo do të interpretohej $\left[(\hat{\beta} - \beta_0)/SE\right]^2 \leq z^2_{\alpha/2}$. Përfundimisht intervali besimit është$\hat{\beta} \pm z_{\alpha/2} * SE$.
Por, karakteristika të tjera mund të kenë rëndësi më të madhe se β, p.sh.përcaktimi I vlerës së $p(x)$ për x të ndryshme . Për $x = x_0$ të fiksuar, $logit[\hat{p}(x_0)] = \hat{\alpha} + \hat{\beta}x_0$ ka një SE që jepet nga rrënja katrore e llogaritur nga vlerat e vleresuara të parametrave:

$$var(\hat{\alpha} + \hat{\beta}x_0) = var(\hat{\alpha}) + x_0^2 var(\hat{\beta}) + 2x_0 cov(\hat{\alpha}, \hat{\beta})$$

Një interval besueshmërie 95% për $logit[p(x_0)$ është $(\hat{\alpha} + \hat{\beta}x_0) \pm 1.96SE$. Por, këto intervale vlejne nëse vëllimi I zgjedhjes është I madh. Duke zëvendësuar çdo skaj në transformimin e anasjelltë,

$$p(x_0) = \exp(logit)\,/[1 + \exp\,(logit)] \quad (11)$$

merret një inteval korespondues $p(x_0)$.
Çdo metodë e paraqitur mund të sigurojë interval besueshmërie dhe në teste me vëllim të vogël.

5.2.2 Zbatim 1 (vazhdim)

Tabela 5-2 tregon rezultatet e hipotezës se faktori, me anë të së cilit modelimi po ndërtohet, mund të mos marrë pjesë në modelim. Sic përmendëm, kjo bëhet me ane të

testit Wald, vlerat e vrojtuara te të cilit paraqiten në tabelë. Ndërsa p_vlerat përkatëse, ndihmojne në marrjen e vendimit.
Rezultetet e regresit logjistik, tabela 5.4, tregojne se faktorë të rëndesishëm në këtë modelim janë *'area'* dhe *'duhan_shtëpi'*.
Koeficenti $\beta = 1.028$ pranë *'areas'*, tregon se kur *'area'* rritet me një njësi, pra kur kalohet në zonën e ndotur, dhe vlerat e variablave të tjerë që marrin pjesë në modelim nuk ndryshojnë, mundësitë e kollës rriten, ndaj mundesive për kollë në zonen e kontrollit. Mundesitë për kollë në *zonën e ndotur* janë exp(1.028)=2.798266 e mundesive për kolle në *zonën e kontrollit*.

Në të njëjtin përfundim arrihet edhe duke parë vlerën pranë *'duhan_shtëpi'* 0.122, faktorit tjetër domethënës në këtë modelim, rritja e tij, pra konsumi i duhanit, rrit mundesine e shfaqjes se kollës ndaj rasteve pakollë. Vlera e kësaj rritje mund të vlerësohet me exp(0.122)=1.13, Kështu, në zonën e ndotur mundesia që të shfaqet ky tregues i rëndësishëm i problemeve pneumologjike, është 13% më e lartë se në zonën e kontrollit.

Tabela 5–2 Koeficientët e variablave në ekuacionin e modelimit

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	area	1.028	.326	9.967	1	.002	2.796	1.477	5.293
	pini_duhan	.147	.489	.090	1	.764	1.158	.444	3.018
	duhan_shtepi	.122	.037	10.602	1	.001	1.130	1.050	1.216
	Constant	-2.267	.596	14.491	1	.000	.104		

a. Variable(s) entered on step 1: area, pini_duhan, duhan_shtepi.

5.3 Përputhshmëria e modelit logjistik

Në praktikë, nuk ka garanci që një model i caktuar i regresionit logjik përputhet mirë me të dhënat.
Në rastin e modelit te regresit linear shumëpërmasor, zgjidhja e këtij problemi është relativisht e lehte. Ndërsa këtu nderlikohet.
Në këto raste problemi synohet të zgjidhet duke analizuar përputhshmërinë e modelit logjistik.
Kur variablat pjesëmarrës në modelim janë vetëm kategorikë, përdoret kjo metodë për vlerësimin e mungesës së përputhshmërisë : krahasohen denduritë e vlerave të modeluara, me denduritë e matura, për $y = 0$ dhe $y = 1$. Në çdo konfigurim të x, mund të shumëzohet probabiliteti i vlerësuar i dy rezultateve Y, me numrin e subjekteve në atë bashkësi, duke marrë një vlerësim për densitetin e parashikuar. Këto quhen *vlera të përputhshmerise*.
Test i përputhshmërisë krahason vlerat e vëzhguara, dhe vlerat e përputhshmerisë se parashikuar, duke përdorur testin Pirson X^2. Për një numër të caktuar konfigurimesh, testi I PirsonX^2 ka shpërndarje Hi-katror $X^2(1)$.
Arsyeja për përdorimin e kësaj metode, testit të përshtatshmërisë, kur variablat në modelim jane kategorik , kuptohet se lidhet me faktin, se, nëse do te kishim ndërtuar tabelen e kontigjences për variablat e vazhdueshëm, ajo mund të përmbante vlera

shumë të vogla. Atehere mund të ishte e pamundur krahasimi me anë të testit Hi-katror.
Për më tepër, me rritjen e numrit të variablave që sigurojne modelimin, grupimi i njëkohshëm i vlerave, për çdo kombinim vlerash të variablave nv model, mund të prodhojë një tabelë pasigurie, me një numër të madh qelizash, shumica e të cilave kanë vlerë të vogël.

Sipas **Hosmer dhe Lemeshow** (1980), problemi në këto raste zgjidhet kështu:
Vlerat e vëzhguara dhe të përputhshmërisë, ndahen në grupe sipas probabilitetit të suksesit të vlerësuar, duke përdorur të dhënat origjinale. Rruga që përdoret zakonisht, është ndarja në grupe (p.sh. 10) me vëllim të njëjtë. Çifti i parë i vlerave të vëzhguara dhe vlerave koresponduese të përputhshmërisë, i referohet $n/10$ vëzhgimeve, që kanë probabilitetin e më të lartë të vlerësuar. Çifti tjetër i referohet $n/10$ vëzhgimeve qe sigurojnë, probabilitetet e dyta të vlerësuara, e kështu me rradhë. Çdo grup ka një numër subjektesh të vëzhguar dhe, një vlerë të përputhshmërisë. Vlera e përputhshmërisë për një rezultat është shuma e probabiliteteve të vlerësuara, për këtë rezultat për të gjitha vëzhgimet në atë grup.
Ata propozuan një test Pirson, duke krahasuar vlerat e vëzhguara dhe të përputhshmërisë për ndarjet e lartëpërmendura.
Testin e ndertuan kështu: Shënojmë me y_{ij} rezultatin binary për vëzhgimin j në grupin $i$të ndarjes, $i = 1, \cdots, n_i$. Le të shënojmë $\hat{p}_{ij}$ probabilitetin e përputhshmërisë korespondues për modelin në fjalëme të dhënat e pagrupuara. Testi është

$$\sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \hat{p}_{ij}\right)^2}{\left(\sum_j \hat{p}_{ij}\right)\left[1 - \left(\sum_j \hat{p}_{ij}\right)/n_i\right]}$$

Kur shumë vëzhgime kanë të njëjtin probabilitet të vlerësuar, ka një arbitraritet në formimin e grupeve, dhe software të ndryshëm mund të raportojnë deri diku vlera të ndryshme.
Testi Hosmer-Lemeshow nuk është shumë I fuqishëm (Hosmer, 1997)[27] për të dedektuar raste të veçanta mungese përputhshmërie. Mënyra e krahasimit të modelit të propozuar, me vlerat reale është më e përdorshme shkencërisht. (*)
Nëse p-vlera e afishuar e testit është e madhe , kjo tregon se përshtatshmëria e të dhënave me modelin është e mirë.

5.3.1 Vërejtje

- Në modelin e regresit linear të shumëfishtë, shih 2.2.1" Për cfarë vlen vlerësimi I koeficientit të korelacionit", është treguar se për te kontrolluar nëse vlen apo jo modelimi I propozuar, kontrollohet hipoteza Ho: $\rho_{y.x_1 \ldots x_n} = 0$. Ndaj kryhet testi ANOVA, e kuptohet se kur hipoteza bie poshte, pra për p-vlera shumë të vogla, konkludohet se *modeli është i vlefshëm*.
- Në rastin e regresit logjistik ndodh ndryshe. Aty, me anë të testit me shpërndarje X^2, kontrollohet hipoteza 'modeli logjistik është i vlefshem'. Ndaj

p-vlera te vogla, tregojnë se testi bie poshte pra, nuk kemi model logjistik të vlefshëm.

5.3.2 Zbatim 1 (vazhdim)

Për modelin e mësipërm është bërë kontrolli i përshtatshmërisë së modelit. Kërkohet të kontrollohet hipoteza:

'ky model logjistik është i vlefshëm',

me hipoteze alternative:

'modeli në fjalë nuk ia vlen të shqyrtohet'.

Afishimet e Tabelës 5-3 japin përgjigjen e hipotezës . Në tabelen e afishuar lexohen vlera e testit X^2, dhe p-vlera e tij.

Tabela 5–3 Kontrolli i vlefshmërisë së modelit

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.205	4	.184

P-vlera 0.184, jo signifikative, tregon se modeli është i vlefshëm.

Pas modelimit afishohet dhe Tabela 5-4 që përmban vlerat e Cox & Snell R Square dhe Nagelkerke R Square. Këto vlera të R^2 s'janë të njëjta, se ato jane llogaritur ne shkalle te ndryshme. Vlera e Cox & Snell R Square, sipas statisticienëve luhatet deri në 0.7, ndërsa Nagelkerke R Square I merr vlerat si zakonisht deri në 1, ndaj zakonisht në interpretim i referohemi këtij vlerësimi. Ky numer jep njëfarë informacioni për perafrimin e modelit.

Tabela 5–4 Informacione për modelimin

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	266.107[a]	.157	.209

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

5.4 Modelimi i vlerave të parametrit të parashikuar

Qëllimi i modelit logjistik është të arrijme, që duke njohur vlerat e variablave që marrin pjesë në modelim të mund të parashikojmë vlerën e variablit të varur. Në programet statistikore, krahas përfundimeve të mësipërmë, në tabelat e data-ve afishohen edhe dy variabla të rinj, vlera e parashikuar e probabilitetit dhe 'grup ku bën pjesë', që në fakt është vlera e variablit që ne synojmë të parashikojme përmes modelimit.Duke lexuar këtë vlerë mund të konkludojmë, ku bën pjesë individi.

Le të shpjegojmë mënyrën se si arrihet të behet kjo. Proçesi kalon në disa etapa:

- Duke shfrytëzuar koeficientët e pastandartizuara të tabeles dhe konstanten e fituar nga analiza e regresit logjistik binar, arrihet te modelohet variabli *predicted_logit_probability*.

- Së dyti, radhiten vlerat e variablit ne fjalë.
  Krijohet nje variabël i ri index dhe me ndihmën e tij dhe të vlerave të radhitura të *predicted_logit_probability,* nëse behet paraqitja grafike, qoftë dhe e tipit scatter, mund të shihet qarte se ky variabel i ri i modeluar me ndihmën e e koeficientëve të tabelës ka formë S, pra është një vijë sigmoidale[28].
- Janë pikërisht vlerat e këtij variabli *predicted_logit_probability* që transformohen në vlerat e variablit *predic_probability.*
- Etapa e kater mbyll procesin e parashikimit, bën të mundur që me anë të një transformimi, vlerat e *predicted_logit_probability*, të kthehen në vlerat e duhura që përcaktojnë se ku bën pjesë individi. Zbatimi i mëposhtëm paraqet saktë këto procese.

5.4.1 Zbatim 1 (vazhdim)

Në zbatimin e paraqitur mësipër,do të tregohet si mund të arrihet, punë që sot bëhet nga softet statistikore, të gjenerohen vlerat e parashikuara.
Së pari, me ndihmën e sintaxes, duke shfrytëzuar konstanten e koeficientët e pastandartizuara të fituar nga analiza e regresit logjistik ndertoj funksionin

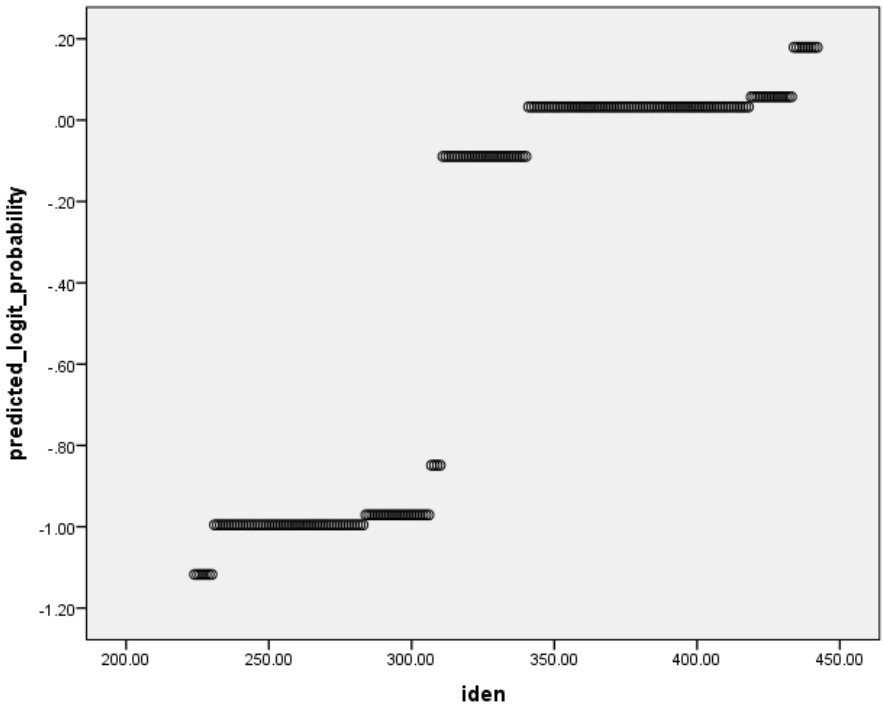
Compute predicted_logit_probability=1.02810746*area+.14654518*pini_duhan
+.12196304*duhan_shtepi -2.26733415

Duhet patur parasysh se këto nuk jane vlerat e parashikuara të probabiliteteve, këto jane vlerat e logit të tyre, pra

$$F(w)=\frac{e^w}{(1+e^w)}$$

Së dyti, radhis vlerat e variablit *predicted_logit_probability*, dhe me ndihmën e variablit index, të krijuar nga ne bëj paraqitjen e tyre me nje grafik scatter që është paraqitur në Figura 5-2

Figura 5–2 Grafiku i predicted_logit_probability



Sa më shumë grafiku ka formën e S aq më i mirë del parashikimi. Në fakt në rastin tonë të modelimit, raporti i numrit të personave me kollë (të koduar 1), me ato pakollë, (të koduar 0)
ishte afërsisht i njëjtë. Nëse do të ishtë shumë i ndryshëm atëhere parashikimi që në fillim pa bërë asnjë modelim do të ishte shumë më i lartë, dhe kur të përmirësohet nga modelimi kuptohet do të sigurohej, së fundmi, vlera shumë më të afërta me ato të matura.

Së treti, duke shfrytëzuar formulën, me ndihmën e sintaksës arrihet të krijohet variabli *predictit_probabilitet*.

```
compute predictit_probabilitet=2.7182818281828**   predicted_logit_probability
                                /(1+2.7182818281828)** predicted_logit_probability
```

Së fundmi, behet transformimi i *predictit_probabilitet* në variabël binar. Transformimi më i llogjikshëm është, që në rastet që ky variabël merr vlera më të vogla 0.5, atëhere vlera parashikohet 0, ndërsa në rastet që ajo është më e madhe se 0.5, vlera e parashikuar e variablit të varur të jetë 1.
Shpesh herë propozohet se si vlerë kufiri përcaktuese, ndarëse, të merret 0.45. Në tabelën 5-5 paraqitet një pjesë e vlerave origjinale të variablit *'kolle*'. Në kollonën e dytë, janë paraqitur një pjesë e vlerave të variablave të sqaruar mësipër, *predicted_logit_probability*, *predictit_probabilitet* dhe *predicted value.*

Tabela 5–5 Të dhenat e modelimit te regresit logjistik

Vl.ak.tkolle	Index	Pred_logit_pr	Pred_prob	Pred_value
.00	6	.03	.39653	.00
.00	7	.03	.39653	.00
1.00	8	.03	.39653	.00
1.00	9	.03	.39653	.00
.00	10	.03	.39653	.00
1.00	11	.03	.39653	.00
1.00	12	.03	.39653	.00
1.00	13	.03	.39653	.00
.00	14	.03	.39653	.00
.00	15	.03	.39653	.00
1.00	16	.06	.85568	1.00
1.00	17	.06	.85568	1.00
1.00	18	.06	.85568	1.00
1.00	19	.06	.85568	1.00
.00	20	.06	.85568	1.00
1.00	21	.06	.85568	1.00
1.00	22	.06	.85568	1.00
1.00	23	.06	.85568	1.00
1.00	24	.06	.85568	1.00
1.00	25	.06	.85568	1.00
1.00	26	.06	.85568	1.00
1.00	27	.06	.85568	1.00
1.00	28	.06	.85568	1.00
1.00	29	.06	.85568	1.00
1.00	30	.06	.85568	1.00
.00	31	-1.00	.24721	.00
1.00	32	-1.00	.24721	.00
.00	33	-1.00	.24721	.00
.00	34	-1.00	.24721	.00
.00	35	-1.00	.24721	.00
.00	36	-1.00	.24721	.00
.00	37	-1.00	.24721	.00
.00	38	-1.00	.24721	.00
1.00	39	-1.00	.24721	.00
1.00	40	-1.00	.24721	.00
1.00	41	-1.00	.24721	.00
1.00	42	-1.00	.24721	.00
1.00	43	-1.00	.24721	.00
.00	44	-1.00	.24721	.00
.00	45	-1.00	.24721	.00
.00	46	1.00	.24721	.00
.00	47	1.00	.24721	.00
.00	48	1.00	.24721	.00
.00	49	1.00	.24721	.00

.00	50	1.00	.24721	.00
.00	51	1.00	.24721	.00
1.00	52	1.00	.24721	.00
.00	53	1.00	.24721	.00
.00	54	1.00	.24721	.00

Po te shihet me kujdes ka nje numër të madh vlerash 1 që parashikohen në rregull, ndërsa pa modelin në fjalë ato nuk parashikohen dot. asnjëra prej tyre. Megjjithëse duhet thëne se më parë parashikoheshin të gjithë rastet pa kollë.

5.5 Funksionit lidhës 'Probabit'

Modeli i ndërtuar ka formën p(x)=F(x). Duke shfrytëzuar klasën e gjërë të shpërndarjes normale me shumëllojshmëritë e pritjes matematike dhe dispersioneve, mund të gjendet një i tillë qe shkallën e rritjes ta ketë të njejtë me ate të shperndarjes normale të normuar. Por, me ndihmen e funksionit shpërndarës të shpërndarjes normale θ, lidhja e mësipërme mund të shkruhet në formën:

$$p(x)= \theta\ (\beta_0 + \beta x \ \ )$$

Meqë θështë funksion monoton rritës, funksioni I anasjelltë θ^{-1}ekziston[29], ndaj barazimi (6) merr formën:

$$\theta^{-1}\big(p(x)\big) = \beta_0 + \beta x \quad (7)$$

Në këtë rast funksioni lidhës është θ^{-1} i cili pasqyron zonën]0;1[të probabiliteteve në
(-∞; +∞), zona ku marrin vlera variablat pjesëmarrës të modelit të regresit. Modeli (7) quhet model *probit*.

5.5.1 Vërejtje:

- Softet e posacme kompjuterike ofrojnë mundësinë e përdoruesit në zgjedhjen e funksionit lidhes. Pra, mund të zgjidhet funksioni *'logit*' I trajtuar më parë, ose funksioni ne fjalë, *'probit*'.
- Softet ofrojnë dhe mundësinë e funksioneve të tjera lidhës.
- Funksinet e tjera nga 'logit' , zakonisht, përdoren nëse kushti i koeficientëve të njëjtë, për të gjitha kategoritë e variablit të varur, nuk plotesohet.

6 Ndërtimi e analiza e një modeli probabilitaro-statistikor për studimin e efektit të ndotjes

Oksidet e sulfurit(SO_x) jane molekula të squfurit dhe oksigjenit. Dioksidi i sulfurit është forma me e zakonshme që gjendet në shtresat e ulta të atmosferës. Ai, në sasitë 1,000 deri në 3,000 mikron per metër kub(μg/m3), është pa arome dhe shije. Ndërsa në përqëndrimet 10,000 μg/m3, ka një aromë të fortë jo të këndshme[1].
Ky gaz ndodhet në sasira më të medha afer zonave industriale, e vecanërisht atyre naftëmbajtëse. Në vendin tonë, fusha e Kucovës dhe Marinëz janë zonat më njohura naftëmbajtëse. Thuajse çdo familje ka brenda truallit të vet, një pus nafte ende në shfrytëzim madje në të shumicën e rasteve edhe dy puse . Banorët ankohen, se këto puse rrjedhin pa pushim dhe ç'ka është më kryesorja ata janë me avari. Amortizimi i grupeve të grumbullimit të naftës, si rrezultat i korrozionit ka sjellë rredhjen në bazën e rezervuarve, në kolektor, saraqineska, e këto rrjedhje pa dyshim shkaktojnë ndotje të mjedisit në tërësi
Burimi i këtij gazi është ndotja jo vetëm nga zjerrja por dhe, nga përpunimi që i bëhet nënprodukteve të tij. Mbetjet e naftës derdhen në rezervuar dhe ndotin ujin e vaditjes duke e nxjerrë jashtë përdorimit, ujë që banorët e këtij fshati dhe të fshatrave fqinjë e përdorin për vaditjen e bimëve bujqësore gjatë verës. Gjithashtu dhe shkalla e lartë e amortizimit të tubacioneve pus- grup dhe grup-impiant, dekantimi si rezultat i kohës së gjatë të shfrytëzimit, korrozioni i shkaktuar nga nafta me sasi squfuri, sjellin përsëri rrjedhje të sasirave të konsiderueshme, përzierje naftë-ujë në mjedis

Nisur nga problemet në fjalë, jeta e banorëve rrezikohet seriozisht. Për këtë nga Instituti i Shëndetit Publik dhe Ministria e Mjedisit merr të dhëna për treguesit kryesor të gjendjes së njerëzve dhe këto të dhëna analizohen, me qellim marrjen e masave të nevojshme nga organet kompetente për zbatimin e kushteve teknike, në shfrytëzimin e puseve të naftës, që ndodhen pranë shtëpive, si psh, përdorimi i filtrave pastrues dhe monitorizimi i ndotjes. Në këtë punim janë shfrytëzuar të dhënat e marra nga këto institucione, për pesë fshatra të zonës Belinë-Sheqishtë, Zharrë, Marinëz, Patos e Kurjan, ne jug të Shqipërise, në Fier, Figura 1. Fshatrat ndodhen pranë njëri tjetrit.

Në baze të të dhënave të Institutit të Gjeoshkencës u bë paraqitja e vendodhjes së puseve në zonë, Figura 2. Nga vëzhgimi i hollësishëm i pozicionimit të tyre kuptohet, pse Kurjan konsiderohet zonë e paster ndërsa fshatrat e tjerë jo. Ato jane te pasura me puse e janë zona të përpunimit të naftës.

Figura 6-1 Pozicioni i zonë së studimit

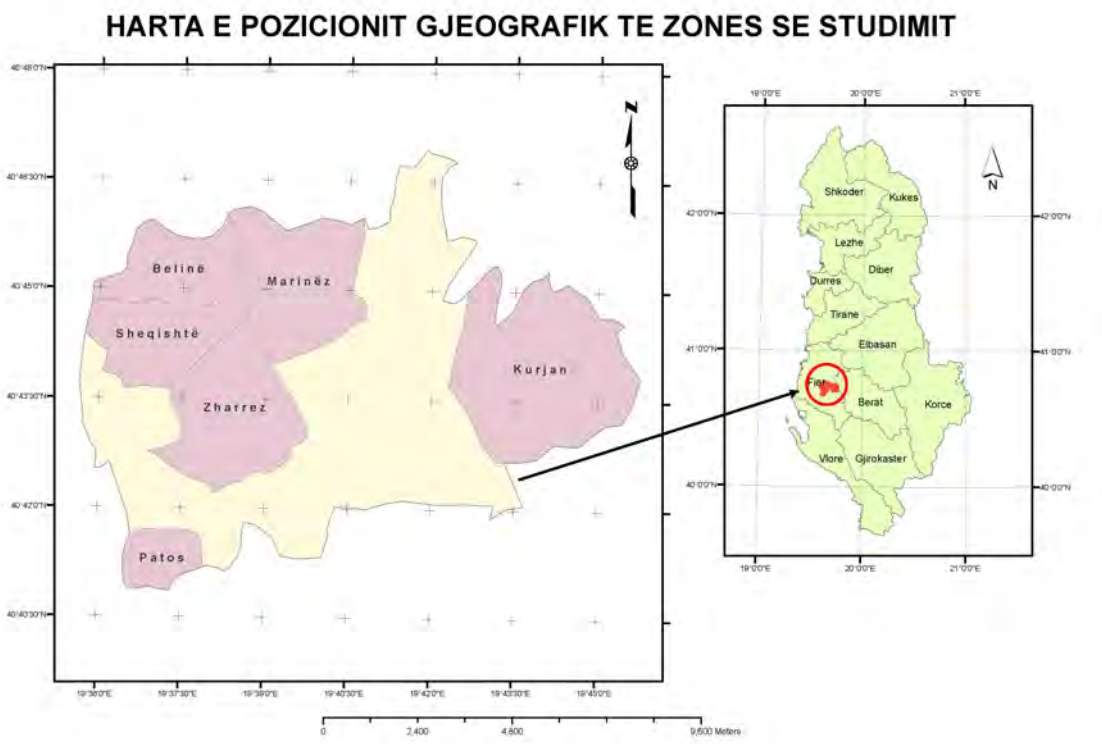
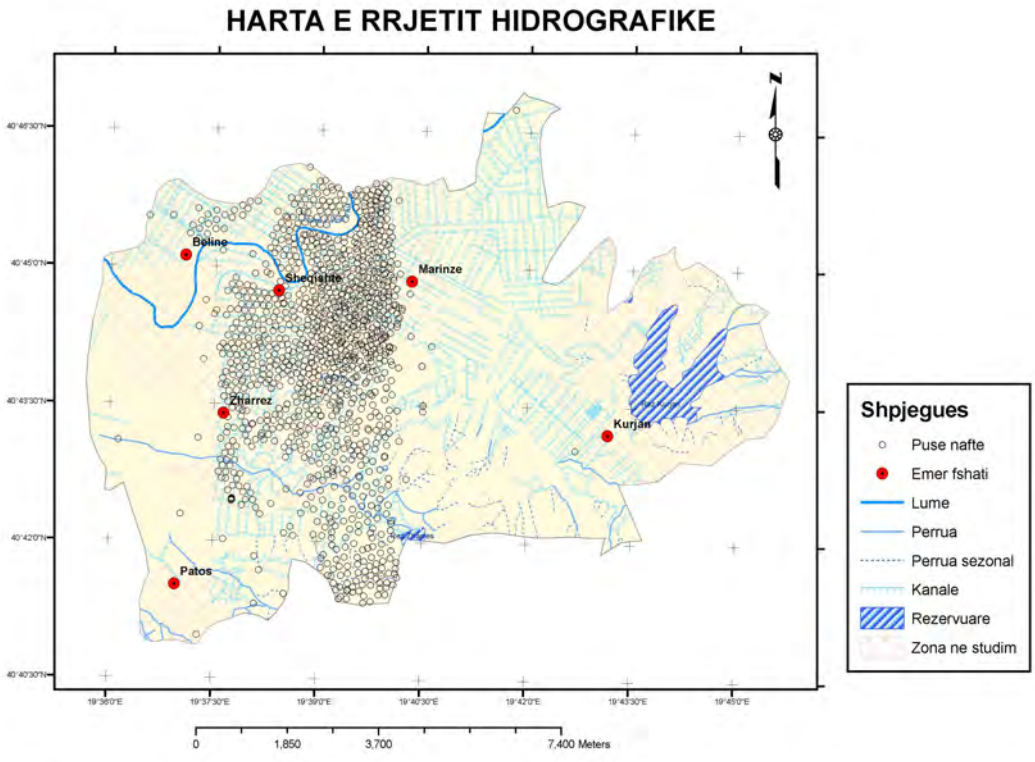


Figura 6–1 Pozicioni i puseve tnaftës



6.1 Ndikimi i nivelit te ndotjes ne problemet shendetësore të përgjithshme, dermatologjike dhe ato pneumologjike. (Problemi 1)

Jane marre të dhëna nga 171 individë, banues te tri zonave, dy të konsideruar industriale, Marinëz dhe Sheqishtë-Belinë, dhe nje e konsideruar e paster, Kurjan. Të intervistuarve i jane berë dhe 13 pyetje te ndryshme. Pyetjet ndahen ne tri kategori: pyetje te *gjendjes së përgjithshme shëndetësore* (3 pyetje), pyetje për *problemet dermatologjike* (5 pyetje) si dhe pyetje për *problemet pneumologjike* (5 pyetje).
Pyetje të gjendjes së përgjithshme janë: *marrje mendje* (p33), *nauze apo të vjella* (p34) , *dhimbje koke* (p35).
Pyetje për problemet dermatologjike janë: *lunga në lëkurë* (p28), *kruajtje në lëkurë* (p29), *shpërthim puçrash* (p30), *skuqje të lëkurë* (p31), dhe *pezmatim të lekurës* (p32).
Pyetjet për problemet pneumologjike janë: sy të përlotur(p23), *kruarje hundësh*(p24), *tharje gryke*(p25), *gëlbaze* dhe *kollë.*
Nga kodimi i pergjigjeve të të intervistuesve janë përcaktuar disa variabla, që siç shihet, duhet të konsiderohen si variabla të varur. Pra, variabla të varur janë 13.
Shtohet problemii ndikimit në këto variabla të fshatit, pra të faktit që kemi të bëjme me të dhena të marra nga nje fshat industrial apo një fshat joindustrial.
Në fillim do të trajtohen te tri llojet e pyetjeve si të dhëna të problemit *'shenja, efekte'*, më pas do të shihen si variabla të pavarur ne analizën MANOVA.
Për përdorimin e metodës MANOVA kërkohet

Tabela 6–1 Matrica e kovariancës

Box's Test of Equality of Covariance Matrices[a]

Box's M	.000
F	.000
df1	91
df2	35214.891
Sig.	1.000

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + fshatrat

plotësimin e disa kushte. Një kusht, tej mase i rëndesishëm, që duhet te plotësohet, është kushti I varincës së barabarte mes tri grupeve të të dhënave të krijuara në rastin tonë nga ndarja e të 13 te dhënave të ndryshme, sipas fshatrave ku individi është banor.
Në rastin tonë ku kusht plotësohet. Kjo mund të shihet nga rezultati i vlerës së statistikës Fisher dhr p- vlerës koresponduese. Me gjithatë duhet theksuar se hipoteza në fjalë është e vertetë dhe do te pranohet si e tillë edhe për p-vlera >0.005. Konkretisht, në rastin tonë, pasi të dhenat ndahen në bazë të fshatit , banor i të cilit është individi, varesia e 13 variablave shprehet nga

$$\text{cov}(\mathbf{Y_i}, \mathbf{Y_j}) = \sigma_{ij}\,\mathbf{I} \qquad i,j = 1,2\ldots,p$$

ku σ_{ij} është kovarianca mes variablit të i-të dhe të j-të, e llogaritur janë paraqitur në Tabela 6-2.

Tabela 6–2 Kovarianca mes variablave

Inter-Item Covariance Matrix

fshatrat		sy t perlotur	kruarje hundesh	Tharje gryke	gelbaze	kolle	Lunga ne lekure	kruarje ne lekure	Shperthim Pucrrash apo flluskash	skuqje xehtesi ne fytyre	pezmatime te lekures	marramendje	nauze, te vjella	dhimbje koke te forta
Vlosh	sy te perlotur	.207	.005	.058	.037	.000	-.005	-.005	-.005	-.005	-.005	.078	.037	.111
	kruarje hundesh	.005	.088	.014	.012	-.005	.015	.015	-.002	-.002	-.002	.037	.012	.021
	tharje gryke	.058	.014	.157	.025	.038	-.003	-.003	-.003	.013	-.003	.058	.073	.090
	gelbaze	.037	.012	.025	.166	.082	-.003	-.003	-.003	-.003	-.003	.021	.054	.021
	kolle	.000	-.005	.038	.082	.176	-.004	-.004	.013	-.004	.013	.000	.050	.032
	lunga ne lekure	-.005	.015	-.003	-.003	-.004	.016	.016	.000	.000	.000	.012	-.003	-.005
	kruarje ne lekure	-.005	.015	-.003	-.003	-.004	.016	.016	.000	.000	.000	.012	-.003	-.005
	shperthim pucrrash apo flluskash	-.005	-.002	-.003	-.003	.013	.000	.000	.016	.000	.016	.012	-.003	.012
	skuqje nxehtesi ne fytyre	-.005	-.002	.013	-.003	-.004	.000	.000	.000	.016	.000	.012	-.003	-.005
	pezmatime te lekures	-.005	-.002	-.003	-.003	.013	.000	.000	.016	.000	.016	.012	-.003	.012
	marramendje	.078	.037	.058	.021	.000	.012	.012	.012	.012	.012	.207	.037	.078
	nauze, te vjella	.037	.012	.073	.054	.050	-.003	-.003	-.003	-.003	-.003	.037	.166	.085
	dhimbje koke te forta	.111	.021	.090	.021	.032	-.005	-.005	.012	-.005	.012	.078	.085	.207
Sheqishte-Beline	sy te perlotur	.238	.062	.020	.059	-.007	.044	.079	.058	.042	.043	.020	.085	.008

	kruarje hundesh	.062	.220	.095	.029	-.022	.022	.049	.048	.012	.107	-.037	.042	.042
	tharje gryke	.020	.095	.212	.055	.051	.050	.022	.043	.034	.076	.024	.098	.074
	gelbaze	.059	.029	.055	.252	.067	.019	.004	-.002	.050	.029	-.040	.034	-.021
	kolle	-.007	-.022	.051	.067	.254	-.025	.005	-.024	.096	-.003	-.006	.001	.007
	lunga ne lekure	.044	.022	.050	.019	-.025	.142	.088	.116	-.047	.060	-.044	.016	.022
	kruarje ne lekure	.079	.049	.022	.004	.005	.088	.196	.116	.008	.087	-.016	.018	.034
	shperthim pucrrash apo flluskash	.058	.048	.043	-.002	-.024	.116	.116	.165	-.032	.086	-.033	.032	.027
	skuqje nxehtesi ne fytyre	.042	.012	.034	.050	.096	-.047	.008	-.032	.242	.030	.015	.039	.024
	pezmatime te lekures	.043	.107	.076	.029	-.003	.060	.087	.086	.030	.220	-.075	.023	.042
	marramendje	.020	-.037	.024	-.040	-.006	-.044	-.016	-.033	.015	-.075	.212	.022	.055
	nauze, te vjella	.085	.042	.098	.034	.001	.016	.018	.032	.039	.023	.022	.249	.019
	dhimbje koke te forta	.008	.042	.074	-.021	.007	.022	.034	.027	.024	.042	.055	.019	.115
Marinez	sy te perlotur	.238	.062	.020	.059	-.007	.044	.079	.058	.042	.043	.020	.085	.008
	kruarje hundesh	.062	.220	.095	.029	-.022	.022	.049	.048	.012	.107	-.037	.042	.042
	tharje gryke	.020	.095	.212	.055	.051	.050	.022	.043	.034	.076	.024	.098	.074
	gelbaze	.059	.029	.055	.252	.067	.019	.004	-.002	.050	.029	-.040	.034	-.021
	kolle	-.007	-.022	.051	.067	.254	-.025	.005	-.024	.096	-.003	-.006	.001	.007
	lunga ne lekure	.044	.022	.050	.019	-.025	.142	.088	.116	-.047	.060	-.044	.016	.022
	kruarje ne lekure	.079	.049	.022	.004	.005	.088	.196	.116	.008	.087	-.016	.018	.034
	shperthim pucrrash apo flluskash	.058	.048	.043	-.002	-.024	.116	.116	.165	-.032	.086	-.033	.032	.027

	skuqje													
	nxehtesi ne	.042	.012	.034	.050	.096	-.047	.008	-.032	.242	.030	.015	.039	.024
	fytyre													
	pezmatime					-								
	te lekures	.043	.107	.076	.029	.003	.060	.087	.086	.030	.220	-.075	.023	.042
	marramendj					-								
	e	.020	-.037	.024	-.040	.006	-.044	-.016	-.033	.015	-.075	.212	.022	.055
	nauze, te													
	vjella	.085	.042	.098	.034	.001	.016	.018	.032	.039	.023	.022	.249	.019
	dhimbje													
	koke te forta	.008	.042	.074	-.021	.007	.022	.034	.027	.024	.042	.055	.019	.115

Afishimet e marra nga metoda e Multivariancës janë paraqitur në Tabelën 6-3.
Vlera e statistikës Fisher, në rastet ku nuk i është kushtuar kujdes normalizimit të të dhënave, lexohet nga afishimi i Pillai's Trace, është 4.654 dhe i korespondon p- vlera 1.2 E-12. Kjo tregon se ka efekt variabli i pavarur , ne rastin tonë *'fshatrat'*, në vlerat e gjithe variablave të varura, kur ato të 13 shihen si një grup i vetëm. Madje, p-vlera tej mase e vogël, percakton dhe shkallën e lartë të ndryshimit të 'efekteve' sipas fshatrave. Duhet të theksohet se në rastin tone efekti është signifikativ sipas të gjitha modeleve të ofruara, madje edhe sipas testit Roy[30]. Ky test përdoret pasi të jenë plotësuar shumë kushte. Psh, kërkon jo thjeshtë kontrollin e kufijve të disa parametrave që na japin informacion për normalizimin, por patjeter plotësim të plotë të kushteve të normalizimit.

Tabela 6–3 Testi i Multivariancës

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.992	1447.439[b]	13.000	156.000	.000	.992
	Wilks' Lambda	.008	1447.439[b]	13.000	156.000	.000	.992
	Hotelling's Trace	120.620	1447.439[b]	13.000	156.000	.000	.992
	Roy's Largest Root	120.620	1447.439[b]	13.000	156.000	.000	.992
fshatrat	Pillai's Trace	.556	4.654	26.000	314.000	.000	.278
	Wilks' Lambda	.444	6.015[b]	26.000	312.000	.000	.334
	Hotelling's Trace	1.254	7.474	26.000	310.000	.000	.385
	Roy's Largest Root	1.254	15.141[c]	13.000	157.000	.000	.556

a. Design: Intercept + fshatrat

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Madje në kollonën e fundit lexohet, se sipas testit me të dobët Pillai, 30% e ndryshimeve në gjendjen shëndetësore të njerëzve shpjegohen si rrezultat I tri nivele të ndryshme të variablit 'fshatrat'. Ndërsa sipas testit të fuqishëm Roy, kushtet e të cilit në rastin tonë plotësohen, jemi në kushtet e variancës së barabartë mes grupeve, më shumë se 50% e ndryshimeve në gjendjen e përgjithshme varen nga zona e banimit.
Konkluzioni në të cilin është arritur, duhet të interpretohet drejt, pra ka nje ndikim te variablit '*fshatrat'* tek ndonjë nënbashkësi e variablave të trajtuar deri tani si një grup.
Procesi i MANOVA-s vazhdon me etapën e dytë, atë të kryerjes së disa testeve 'univarate' ndaj secilit prej13 pyetjeve, pra kryhem ANOVA per çdo pyetje. Afishimet tregojne per ndikim të variablit *fshatrat* tek secili nga 13 pyetjet e studiuara. (Tabela 6-4)
Shihet se ndikimi më i vogel I *'fshatrat*' është tek *kruarja e hundeve* dhe *kolle*. Në këtë përfundim arrihet duke u mbështetur tek vlerat e F-testit dhe afishimet koresponduese të p-vlerave.
Duhet theksuar se nuk vërehet ndikim shumë i ndjeshëm i *fshatrat* nga ku është marrë zgjedhja në *'lunga në lëkurë'.*

Tabela 6–4 ANOVA për 13 variablat e varur

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
fshatrat	sy te perlotur	4.706	2	2.353	10.392	.000	.110
	kruarje hundesh	1.918	2	.959	5.610	.004	.063
	tharje gryke	10.481	2	5.240	27.313	.000	.245
	gelbaze	4.852	2	2.426	11.020	.000	.116
	kolle	2.674	2	1.337	5.935	.003	.066
	lunga ne lekure	.905	2	.452	4.755	.010	.054
	kruarje ne lekure	2.357	2	1.179	9.113	.000	.098
	shperthim pucrrash apo filluskash	1.404	2	.702	6.373	.002	.071
	skuqje nxehtesi ne fytyre	14.098	2	7.049	44.434	.000	.346
	pezmatime te lekures	3.556	2	1.778	12.302	.000	.128
	marramendje	6.952	2	3.476	16.507	.000	.164
	nauze, te vjella	5.380	2	2.690	12.306	.000	.128
	dhimbje koke te forta	13.601	2	6.800	45.622	.000	.352

Tabela 6–5 t testet për secilin prej 13 variablave të varur

Dependent Variable	(I) fshatrat	(J) fshatrat	Mean Difference (I-J)	Std. Error	Sig.
sy te perlotur	Kurjan	Sheqishte-Beline	.3439*	.08825	.000
		Marinez	.3439*	.08825	.000
	Sheqishte-Beline	Kurjan	-.3439*	.08825	.000
		Marinez	.0000	.09158	1.000
	Marinez	Kurjan	-.3439*	.08825	.000
		Sheqishte-Beline	.0000	.09158	1.000
kruarje hundesh	Kurjan	Sheqishte-Beline	.2196*	.07668	.005
		Marinez	.2196*	.07668	.005
	Sheqishte-Beline	Kurjan	-.2196*	.07668	.005
		Marinez	.0000	.07958	1.000
	Marinez	Kurjan	-.2196*	.07668	.005
		Sheqishte-Beline	.0000	.07958	1.000
tharje gryke	Kurjan	Sheqishte-Beline	.5132*	.08123	.000
		Marinez	.5132*	.08123	.000
	Sheqishte-Beline	Kurjan	-.5132*	.08123	.000
		Marinez	.0000	.08430	1.000
	Marinez	Kurjan	-.5132*	.08123	.000
		Sheqishte-Beline	.0000	.08430	1.000
gelbaze	Kurjan	Sheqishte-Beline	.3492*	.08701	.000
		Marinez	.3492*	.08701	.000
	Sheqishte-Beline	Kurjan	-.3492*	.08701	.000
		Marinez	.0000	.09030	1.000
	Marinez	Kurjan	-.3492*	.08701	.000
		Sheqishte-Beline	.0000	.09030	1.000
kolle	Kurjan	Sheqishte-Beline	.2593*	.08803	.004
		Marinez	.2593*	.08803	.004
	Sheqishte-Beline	Kurjan	-.2593*	.08803	.004
		Marinez	.0000	.09135	1.000
	Marinez	Kurjan	-.2593*	.08803	.004
		Sheqishte-Beline	.0000	.09135	1.000
lunga ne lekure	Kurjan	Sheqishte-Beline	.1508*	.05720	.009
		Marinez	.1508*	.05720	.009
	Sheqishte-Beline	Kurjan	-.1508*	.05720	.009
		Marinez	.0000	.05936	1.000

	Marinez	Kurjan	-.1508*	.05720	.009
		Sheqishte-Beline	.0000	.05936	1.000
	Kurjan	Sheqishte-Beline	.2434*	.06669	.000
		Marinez	.2434*	.06669	.000
kruarje ne lekure	Sheqishte-Beline	Kurjan	-.2434*	.06669	.000
		Marinez	.0000	.06921	1.000
	Marinez	Kurjan	-.2434*	.06669	.000
		Sheqishte-Beline	.0000	.06921	1.000
	Kurjan	Sheqishte-Beline	.1878*	.06154	.003
		Marinez	.1878*	.06154	.003
shperthim puçrrash apo flluskash	Sheqishte-Beline	Kurjan	-.1878*	.06154	.003
		Marinez	.0000	.06387	1.000
	Marinez	Kurjan	-.1878*	.06154	.003
		Sheqishte-Beline	.0000	.06387	1.000
	Kurjan	Sheqishte-Beline	.5952*	.07386	.000
		Marinez	.5952*	.07386	.000
skuqje nxehtesi ne fytyre	Sheqishte-Beline	Kurjan	-.5952*	.07386	.000
		Marinez	.0000	.07665	1.000
	Marinez	Kurjan	-.5952*	.07386	.000
		Sheqishte-Beline	.0000	.07665	1.000
	Kurjan	Sheqishte-Beline	.2989*	.07050	.000
		Marinez	.2989*	.07050	.000
pezmatime te lekures	Sheqishte-Beline	Kurjan	-.2989*	.07050	.000
		Marinez	.0000	.07316	1.000
	Marinez	Kurjan	-.2989*	.07050	.000
		Sheqishte-Beline	.0000	.07316	1.000
	Kurjan	Sheqishte-Beline	.4180*	.08510	.000
		Marinez	.4180*	.08510	.000
marramendje	Sheqishte-Beline	Kurjan	-.4180*	.08510	.000
		Marinez	.0000	.08831	1.000
	Marinez	Kurjan	-.4180*	.08510	.000
		Sheqishte-Beline	.0000	.08831	1.000
	Kurjan	Sheqishte-Beline	.3677*	.08671	.000
		Marinez	.3677*	.08671	.000
nauze, te vjella	Sheqishte-Beline	Kurjan	-.3677*	.08671	.000
		Marinez	.0000	.08998	1.000
	Marinez	Kurjan	-.3677*	.08671	.000
		Sheqishte-Beline	.0000	.08998	1.000
	Kurjan	Sheqishte-Beline	.5847*	.07160	.000

		Marinez	.5847*	.07160	.000
dhimbje koke te forta	Sheqishte-Beline	Kurjan	-.5847*	.07160	.000
		Marinez	.0000	.07430	1.000
	Marinez	Kurjan	-.5847*	.07160	.000
		Sheqishte-Beline	.0000	.07430	1.000
	Kurjan	Sheqishte-Beline	.5847*	.07160	.000
		Marinez	.5847*	.07160	.000

Duke u perqendruar në vlerat e t-testeve te afishuara mund të arrihet në perfundimin :

- Ka ndryshim signifikativ midis fshatrave Kurjan dhe Sheqishtë- Beline dhe Kurjan e Marinez për çdo pyetje te bëre. Mesatarja e variablave, të përcaktuar nga përgjigjet e çdo pyetje të bërë, është statistikisht më e larte në Kurjan. Duke vërejtur mënyrën e kodimit të përgjigjeve në fjalë, konkludohet se ka ndikim negativ të zonës industriale në tërë parametrat e përcaktuara si percaktues të gjendjes së përgjithshme, të gjendjes pneumologjike dhe asaj dermatologjike.
- Nuk vihet re ndryshim midis dy fshatrave që ndodhen në zonat industriale.

Por metoda e paraqitur deri tani nuk është gjë tjetër veçse metoda me e zakonshme e Manova-s.

Vlera me e madhe e metodës MANOVA verehet në mënyrën tjetër që mund të kryhet kjo metode, në **analizën e komponenteve kryesore**. Përcaktimi i ketyre komponentëve kryesorë është një përcaktim i variablave të rinj që janë kombinim linear i 13 variablave te varur ne studim,e shpesh për këtë arsye njihen me emrin përcaktimi i supervariablave. Këto variabla të rinj, supervariabla[15], të pavarur synojnë të paraqesin të njëjtat matje apo informacione të marra nga 13 variablat në studim.

MANOVA për analizën e komponentëve kryesorë do te ekzekutohet pas shkrimit të kodit në editorin e sintaksës. Editori është emertuar ' *pyetje'* dhe permban kodin e meposhtëm:

```
MANOVA p23 p24 p25 gelbaze kolle p28 p29 p30 p31 p32 p33 p34 p35 BY
fshatrat(1,3)
/DISCRIM=STAN RAW CORR
/PRINT =SIGNIF( MULTIV,UNIV,EIGEN,DIMENR)/DESIGN.
```

Pjesa e pare e afishimeve, Tabela 6-4, tregon per ndikim te variablit '*fshatrat*' ne tërë keto variabla te varur. Këtë informacion e morëm edhe nga ekzekutimi i metodes MANOVA sipas variantit të parë. Afishimet e ndikimit verehet se janë të njëjta me ato të metodës së parë. Madje po te krahasohen me kujdes, vlerat e testit Pillais është e njëjtë , .55629, dhe po ashtu vlerat e testeve të tjera.

Tabela 6–6 Analiza e ndikimit te variablit te pavarur tek variablat e varur në shqyrtim

```
EFFECT .. fshatrat
 Multivariate Tests of Significance (S = 2, M = 5 , N = 77 )

 Test Name            Value        Approx. F       Hypoth. DF         Error DF     Sig. of F

 Pillais                    .55629            4.65353              26.00              314.00
.000
 Hotellings               1.25375            7.47425              26.00              310.00
.000
 Ëilks                      .44371            6.01498              26.00              312.00
.000
 Roys                    .55629
 Note.. F statistic for WILKS' Lambda is exact.
```

Ekzekutimi i kësaj analize siguron gjetjen e dy kombinimeve lineare të 13 variablave në studim. Kombinimi i dy këtyre supervariablave është signifikativ. Dihet se keto dy supervariabla të krijuar ruajnë totalin e variances, dhe janë ortogonalë. Por, supervariabli i parë shpjegon gjithë variancen e tri variablave te pavarur, që fshihen në 13 variablat e varur.

Tabela 6–7 Vlerat e veta

```
 Root No.        Eigenvalue         Pct.       Cum. Pct.      Canon Cor.

        1               1.25375          100.00000           100.00000
.74585
        2                .00000              .00000           100.00000
.00000
```

Por afishimet e Tabeles 6-7 tregojnë se 0.74585^2= 55.6% të variancës supervariablit shpjegohet nga ndryshimet e vlerave të variablit 'fshati'.
Në tabelen 6-8 afishohen vlerat përcaktuese të supervariablit në fjalë, koefiçientët e ketij supervariabli.

Tabela 6–8 Koefiçientet e variablit kanonik

```
Correlations between DEPENDENT and canonical variables
           Canonical Variable

 Variable                   1

 p23                    -.31412
 p24                    -.23080
 p25                    -.50926
 gelbaze                -.32348
 kolle                  -.23739
 p28                    -.21248
 p29                    -.29417
 p30                    -.24600
 p31                    -.64955
 p32                    -.34177
 p33                    -.39591
```

```
p34                 -.34184
p35                 -.65818
```

Pra, këtë variabël kryesor e përcaktojmë si kombinim linear të koefiçientëve të afishuar dhe të 13 variablave në fjalë. E llogarisim këtë variabel dhe e emërtojmë '*superndikim*', duke shkruar kodin përkates në editorin e sintakses.

```
compute super_ndikim=     -.20647* p23 +.00410*p24-.5260*p25+.54939* gelbaze
                          -.10457* kolle          +.90458*p28  +.00185*p29
                          +.19729*  p30+1.58608*  p31+.28266*  p32  +.64448*
                          p33+.16127* p34+1.20061* p35
```

Tabela 6–9 Analiza deskriptive e *fshatrav*e në '*superndikim*'

Descriptive Statistics

Dependent Variable: super_ndikim

fshatrat	Mean	Std. Deviation	N
Vlosh	8.8516	.70669	63
Sheqishte-Beline	6.7938	1.01302	54
Marinez	6.7938	1.01302	54
Total	7.5519	1.34653	171

Kryhet metoda e ANOVA-s në variablin e ri, për të përcaktuar ndikimin në të fshatrave. Afishimet e vlerave të statistikës descriptive të variablit të ri '*superndikim'*, në të tri nënpopullimet e përcaktuara nga vlerat e variablit '*fshatrat'* janë paraqitur në Tabelen 6-9. Mesataret e nenpopullimeve te superndikimit jane te njejta, më të vogla, për dy fshatrat që ndodhen në zonën industriale, dhe po kështu dhe devijimet standarte. Duke patur parasysh kodimin e 13 variablave (sa më I lartë kodi aq më e mirë gjendja e përgjithshme), themi se ka ndikim negativ ndotja, efekt negativ në gjendjen e përgjithshme të individid.

Afishimet e ANOVA-s për '*superndikim*'in që shpjegon maksimumin e variancës mes individeve,Tabela 6-10, tregon ndikimin e *fshatra*-ve në variablin e ri. Madje tabela ne fjalë tregon se 55 % e ndryshimeve të *superndikimi*-t vijne si rezultat I ndikimit të *fshatra*-ve (vlera e statistikës F është 101.28 ndërsa p-vlera <.001)

Tabela 6–10 Efekti në 'superndikimi' i variablit të pavarur

Tests of Between-Subjects Effects

Dependent Variable: super_ndikim

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	168.492[a]	2	84.246	101.282	.000	.547
Intercept	9516.397	1	9516.397	11440.809	.000	.986
fshatrat	168.492	2	84.246	101.282	.000	.547

Error	139.741	168	.832			
Total	10060.576	171				
Corrected Total	308.234	170				

a. R Squared = .547 (Adjusted R Squared = .541)

Vihet re se:

- I njejti problem i trajtuar me metoden e komponentëve kryesorë, tregon se ndotja ndikon në 55% të gjendjes së përgjithshme (të shprehur nga variabli 'super_ndikim'), ndërsa sipas metodes MANOVA(GLM), ndotja spjegonte 27.8-50% te variances së 13 komponentëve të gjendjes së përgjithshme.
- Zgjidhja e problemit me MANOVA (GLM) është me e ndërlikuar e ka një vëllim të madh analizash, ndërsa 'analiza e komponenteve kryesore' e redukton problemin në ndikimin në një variabel të vetëm, për të cilin mund të vlerësohen qartë dhe vlerat mesatare të tij në tri nivelet e variablit 'fshatrat'.

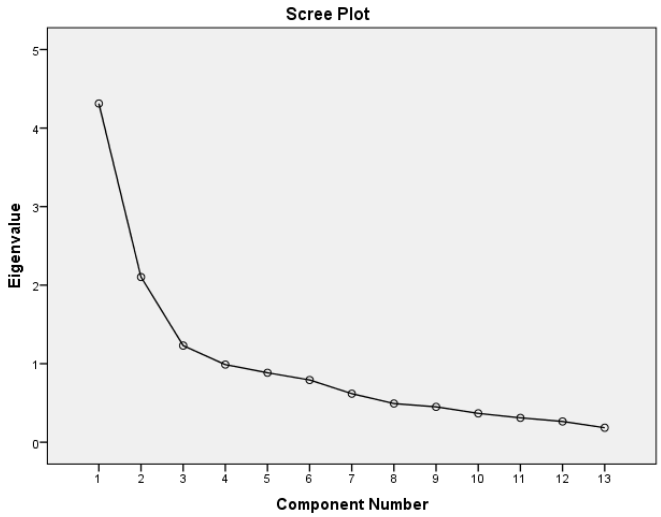
Problemi në fjalë mund të trajtohet edhe me anë të *Metodës se Analizës Faktoriale*. Siç është përmendur në vërejten 1, metoda MANOVA, do të ishte e pakuptimtë të përdorej nësë tërë faktoret, variablat e varur do te ishin te ndryshëm. Këto variabla u tha se duhet t'i perkisnin të paktën dy grupeve të ndryshme. Në rastin tonë ato I perkasin tri grupeve të ndyshme. Tri pyetje kanë të bëjnë me gjendjen e përgjithshme të individit, 5 me problemet e tij dermatologjike dhe 5 me problemet pneumologjike.
Shkrimi i kodit të meposhtëm përdoret për të kryer këtë metodë:

```
FACTOR
  /VARIABLES p23 p24 p25 gelbaze kolle p28 p29 p30 p31 p32 p33 p34
p35
  /MISSING LISTWISE
  /ANALYSIS p23 p24 p25 gelbaze kolle p28 p29 p30 p31 p32 p33 p34 p35
  /PRINT INITIAL SIG EXTRACTION ROTATION
  /PLOT EIGEN
  /CRITERIA FACTORS(3) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION.
```

Afishimet e mëposhtme Figura 2, tregojnë semerren tri komponente kryesore. Kjo sepse, sic shihet edhe nga grafiku, vlerat e veta jane më të larta për 3-4 komponentët e parë. Madje kjo verehet edhe ne afishimet e tabeles së Analizës së Komponenetëve Kryesorë.

Figura 6–3 Plotimi I vlerave të veta

a. Rotation converged in 4 iterations.



Pesë komponente kryesore shpjegojnë më shumë se 73% të variancës së përgjithshme. Në rastin tonë ne po përqëndrohemi në tri komponentët e parë kryesorë qe shpjegojnë rreth 60% të kësaj variance të përgjthshme, sepse siç shihet vetëm tri vlerat e veta te para janë >1.

Në këtë rast do të krijoheshin tri variabla te rinj, variabla që janë *'faktore të 13 variablave'*.

Tabela 6–11 Analiza e Komponentëve Kryesorë

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Comp.	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.313	33.174	33.174	4.313	33.174	33.174	3.091	23.776	23.776
2	2.103	16.179	49.353	2.103	16.179	49.353	2.614	20.111	43.887
3	1.231	9.466	58.820	1.231	9.466	58.820	1.941	14.933	58.820
4	.988	7.602	66.421						
5	.885	6.804	73.225						
6	.793	6.097	79.322						
7	.618	4.754	84.076						
8	.493	3.790	87.866						
9	.450	3.463	91.329						
10	.368	2.828	94.157						
11	.311	2.390	96.547						

12	.264	2.035	98.582						
13	.184	1.418	100.000						

Analiza faktoriale realizon krijimin e 3 variablave të rinj, e në përputhje me tri grupimet e 13 variablave të varur këto variabla po i emërtojme *efektifaktorial1, efektifaktorial2, efektifaktorial3*.
Përqëndrimi në vlerat e afishuara te koeficienteve prane 13 variablave për tre komponentët e rinj *efektifaktorial,* Tabelen 6-12, tregon se komponenti i pare ka koeficientë më të lartë pranë ndikimit ne lekure, ndaj ndryshe efektifaktorial1 mund të emertohet *komponent_dermatologjik*.
Ky variabël justifikon 33% të ndryshimeve të variancës totale. Komponenti I dyte I ka koefiçientët më të lartë pranë treguesve të gjendjes së përgjithshme (*komponent_pergjithshem*), dhe te dy komponentet e rinj të parë sëbashku justifikojnë rreth 50% të variancës së përgjithshme. Ndersa i treti është më I koreluar me pyetje qe kanë të bëjnë me fushën pneumologjike, ndaj *efektifaktorial3* mund të emërtohet *komponent_pneumologjik*.
Ndikimi i 'fshat'-it vërehet në te tri komponentet, dhe ai është signifikativ në te tri komponentet, por ai percakton 44% të ndryshimit të variablit perfaqësues te gjendjes së përgjithshme dhe 35% të variablit të gjendjes se përgjithshme pneumologjike, ku variabli 'kolle' ze vendin kryesor, me koefiçientin me te lartë. Ndikimi I këtij variabli justifikon vetëm 22% të ndryshimeve ne komponentin dermatologjik, që përcaktohet nga lunga, puçra dhe kruarje në lëkurë.

Nëse krahasojme përfundimet me metoden MANOVA (Analiza e Komponenteve Kryesor) dhe Analizën Faktoriale vërehet se ndikimet e fshatit , ndotjes, në variablat e varur janë shume larg 55% të fituar me metoden MANOVA. Por në këtë përfundim mund të na çojë vetëm kjo metodë, se kur krijon komponentët përbërës kjo metodë bazohet në idene e maksimizimit të variances.

Tabela 6–12 Afishimet për të tri faktorët

Rotated Component Matrix[a]

	Component		
	1	2	3
sy te perlotur	.324	.542	.122
kruarje hundesh	.460	.273	.178
tharje gryke	.312	.614	.405
gelbaze	.142	.098	.730
kolle	-.062	.063	.783
lunga ne lekure	.834	.026	-.084
kruarje ne lekure	.773	.165	.034
shperthim puçrrash apo flluskash	.872	.104	-.068
skuqje nxehtesi ne fytyre	.020	.437	.600
pezmatime te lekures	.692	.108	.310

Marramendje	-.168	.830	-.104
nauze, te vjella	.161	.576	.275
dhimbje koke te forta	.256	.770	.167

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax With Kaiser Normalization.

Tabela 6–13 Anova për komponentin pneumologjik

Tests of Between-Subjects Effects

Dependent Variable: Komp_pneumologjik

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	70.889[a]	2	35.444	45.764	0	0.353	91.528	1
Intercept	4715.665	1	4715.665	6088.652	0	0.973	6088.652	1
fshatrat	70.889	2	35.444	45.764	0	0.353	91.528	1
Error	130.116	168	0.775					
Total	5026.33	171						
Corrected Total	201.005	170						

a. R Squared = .353 (Adjusted R Squared = .345)

b. Computed using alpha = .05

Tabela 6–14 Anova për komponentin e përgjithshëm

Tests of Between-Subjects Effects
Dependent Variable: Komp_pergjith

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	155.513[a]	2	77.757	67.716	0	0.446	135.431	1
Intercept	8269.91	1	8269.91	7201.985	0	0.977	7201.985	1
fshatrat	155.513	2	77.757	67.716	0	0.446	135.431	1
Error	192.911	168	1.148					
Total	8828.331	171						
Corrected Total	348.425	170						

a. R Squared = .446 (Adjusted R Squared = .440)

b. Computed using alpha = .05

Tabela 6–15 Anova për komponentin e dermatologjik

Tests of Between-Subjects Effects

Dependent Variable: Komp_dermatologjik

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observe Power[b]
Corrected Model	61.298[a]	2	30.649	24.372	.000	.225	48.745	1.000
Intercept	11116.743	1	11116.743	8840.170	.000	.981	8840.170	1.000
fshatrat	61.298	2	30.649	24.372	.000	.225	48.745	1.000
Error	211.264	168	1.258					
Total	11568.856	171						
Corrected Total	272.562	170						

a. R Squared = .225 (Adjusted R Squared = .216)

b. Computed using alpha = .05

6.2 Ndërtimi I modelit statistikore për problemevet pneumologjike të banorëve të zonës industriale (Problemi 2)

Në pjesën e parë të studimit u përqëndruam te shqetësimet e banorëve të zonës së ndotur. Në etapën e fundit të këtij studimi, pasi përdorëm metodën e Analizës së Komponenteve, u arrit të ndërtohej një 'supervariabël' i problemeve pneumologjike, i cili u emërtua 'komponent pneumatik'. Në këtë pjesë të tretë duam të studiojmë problemin pneumatik, dhe të mund ti pergjigjemi pyetjeve: A ndikon ndotja në këtë probleme? A ndikojne në këto probleme faktorë apo variabla të tjerë ?

6.2.1 Modeli i analizës mbi komponentët e Analizës Faktoriale

Në trajtimin e problemit të ndikimit apo jo të zonës së pneumologji, variabël bazë u përdor variabli 'komponent pneumatik'. Kujtojme se variabli ne fjalë është llogaritur me anë të metodës Analiza Faktoriale sipas formulës

```
compute Komp_pneumologjik= .122* p23+.178*p24+.405*p25+.730* gelbaze+
                           .783* kolle - .084*p28 +.034*p29-.068*
                           p30+.600* p31+.310* p32-.104* p33+.275*
                           p34+.167*p35
```

Kujtojmë, se ndonëse ai është kombinim i përgjigjeve të 13 pyetjeve, koeficientet me te lartë, kur ai modelohet i ka pranë 'kollës' e 'gëlbazës' e 'kruarjes në grykë', ndërsa koeficientët e tjerë jane më të vegjel se 0.4. Për zgjidhjen më të mirë të problemit, u llogarit 'komponenti pneumatik specifik', komponent ky vetëm I faktorëve pneumatik.

```
compute Komp_pneumologjik_specifik= .122* p23+.178*p24+.405*p25+
                                    .730* gelbaze+.783* kolle
```

Duke parë mënyrën se si ai është krijuar, themi se, ky variabël është i vazhdueshëm.

Por, kontrolli i problemit të ndikimit të zonës ne këtë variabël, është i pakuptimtë, pa studimin e faktorëve të tjerë që mund të ndikojnë në të. Ndaj, nëse kërkojme të analizojmë ndikimin apo jo në të disa variablave të tjerë, duhet përdorur metoda e regresit linear të shumëfishtë.
Nga analiza deskriptive e të dhënave u vu re 14.3% e të intervistuarve konsumojnë duhan, pjesa tjetër nuk konsumon.
Ndertimi i modelit linear me variablat *'konsum duhani'* dhe '*area'* japin afishimet e mëposhtme, Tabela 16.
Rezultatet e afishuara tregojne se modeli i regresit të ndërtuar, është i vlefshëm. Hipoteza se dy variablat e modelimit nuk kanë lidhje me 'komponent_pneumatik_specifik', bie poshtë. Vlera e statistikës Fisher 46.465 dhe p-vlera shumë e vogël, Tabela 6-16, tregojnë pikërisht këtë.
E modeli i propozuar, duke u bazuar në verejtjes që njihet me emrin ' rule of thumb', Tabela 16 siguron një korelacion të lartë pozitiv me R=0.599. Ai justifikon rreth 36% të ndryshimeve të variablit në shqyrtim, 'komponent pneumatik specifik ', Tabela 6-16.

Tabela 6–16 Afishimet e regresit linear

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.598[a]	.358	.350	.87515055 31	.358	46.465	2	167	.000

a. Predictors: (Constant), a pini duhan, area

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	6.749	.515		13.100	.000		
	area	-1.322	.139	-.588	-9.475	.000	1.000	1.000
	a pini duhan	.374	.237	.098	1.581	.116	1.000	1.000

a. Dependent Variable: Komp_pneumologjik

Ndërsa, kontrolli i hipotezave për domosdoshmërine e variablave në modelim, me anë të t-testit, tregojnë se variabli *'area*' ka ndikim të rëndësishëm në modelin e propozuar.
Modelimi me regres të shumëfishte i *'komponentin_pneomologjik_specifik'*, duke përfshirë dhe variabla të tjerë si, *'konsumoni_duhan_në_shtëpi*, ' *vitesh_banorë_në_zonë'*, variabël i fundit mund të konsiderohet i vazhdueshëm, e rrit efikasitetin e modelit. Modeli justifikon tani 43% të ndryshimeve të *'komponentin_pneomologjik_ specifik* '

Dihet se minimumi i raportit të vëllimit të zgjedhjes që shfrytëzohet me variablat e pavarur duhet të jetë 5 me 1, ndërsa vlera e raportit të preferuar është 15 me 1. Në rastin në fjalë, ky raport më i larte se 20.

Tabela 6–17 Rregulli i vlerësimit të regresit të shumefishtë linear

Rregulli i njohur i vlerësimit të modelit te regresit[11]

Koeficienti I korelacionit	Interpretimi
.80 në 1.00 (-.80 në –1.00)	Të koreluara shumë fort pozitivisht (negativisht)
.60 në .80 (-.60 në -.80)	Të koreluara fort pozitivisht (negativisht)
.40 në .60 (-.40 në -.60)	Mjaftueshëm të koreluara positivisht (negativisht)
.20 në .40 (-.30 në -.50)	Të koreluara dobët positivisht (negativisht)
.00 në .20 (.00 në -.20)	Aspak të koreluara

Tabela 6–18 Koeficientët e modelit të regresit të shumëfishtë

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.656[a]	.430	.408	.8356693197	.430	20.158	4	107	.000

a. Predictors: (Constant), a pi duhan njeri ne shtepine tuaj, a pini duhan, area, prej sa vitesh banoni ne kete zone

Coefficientsa

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	6.343	.943		6.726	.000		
	area	-1.222	.185	-.562	-6.619	.000	.740	1.350
	a pini duhan	.005	.401	.001	.012	.991	.908	1.101

	prej sa vitesh banoni ne kete zone	.001	.005	.011	.134	.894	.730	1.369
	a pi duhan njeri ne shtepine tuaj	.524	.200	.202	2.613	.010	.896	1.117

a. Dependent Variable: Komp_pneumologjik_specifik

Tabela e afishimeve të testeve, për vlerën apo jo të variablave ne modelim, tregon, se nga kater variablat e shqyrtuar, ndikim ne modelim ka *zona* në të cilen individi është banor, pra fakti që ai jeton në zonën e ndotur apo jo, dhe a *konsumon njeri duhan në shtëpi*.
Vlera e koeficientit është negative prane '*area*' , duke treguar se kur indeksi zonë bëhet 2, pra kur zona ndotet, vlera e variablit *'komponent_ pneumologjik_specifik'* zvogelohet. Duke u perqëndruar tek mënyra e kodimit të të dhënave në bazën origjinale , pergjigjet 2 tregojne se nuk ka shenja te indikatorit pneumologjik në fjalë, ndersa 1 ka shenja të indikatorit. Kështu, variabli 'komponent pneumologjik specifik' zvogelohet, do të thote se ka me shume vlera 1 në përgjigjet për indikuesit pneumatik. Pra, se ka *me shume raste te shfaqjes se indikuesve të problemeve pneumologjike, kur kalohet në zonen e ndotur.*
Ndërsa, vlera e koeficientit pranë komponentes tjeter signifikative *'konsumi duhan në shtëpi'*, është pozitive. Kodimi i variablit *'konsumi duhan në shtëpi'*, me 2 ne rast mos konsumi dhe 1 në rast konsumi, tregon se konsumi i duhanit në shtëpi (zvoglimi me nje njësi e variablit 'konsumi duhan në shtëpi') shoqerohet me zvogëlim të *'komponent pneumologjik specifik'*. Bazuar në arsyetimet e mësipërme në *me shume raste te shfaqjes se indikuesve të problemeve pneumologjike, ndërsa kemi konsum duhani në shtëpi dhe vlerat e faktoreve të tjerë nuk ndryshojnë.*
Duke patur parasysh perfundimet tek 2.4, vlera e koeficientëve të modelit të regresit linear të ndërtuar, tregon shkallën e ndryshimit të *'komponent pneumologjik specifik'* në varësi të variablit, pranë të cilit ai ndodhet, kur komponentet e tjerë ne modelim mbahen konstantë.
S' mund të konkludohet bazuar në madhesinë e këtyre koeficientëve në vlerë absolute, për kontribut më të madh në modelim të një variabli se variabli tjetër.

6.2.2 Përfundime

- Ka më shume raste te shfaqjes se indikuesve të problemeve pneumologjike, kur kalohet në zonen e ndotur.
- Ka me shume raste te shfaqjes se indikuesve të problemeve pneumologjike, ndërsa kemi konsum duhani në shtëpi dhe vlerat e faktoreve të tjerë nuk ndryshojnë.

Analiza e problemeve pneumatike të banorëve, në bazë të një indikatori të vetëm

Me interes është kontrolli i ndikimit në, **indikatorin më të rëndesishëm ne pneumologji, '*kolle*'.** Për të analizuar problemin në fjalë janë shfrytëzuar të dhënat e 426 subjekteve të marra rastësisht nga pesë fshatrat e sipërpërmendura, 196 subjekte janë nga zona e pastër, zona e kontrollit, pjesa tjetër 230 janë marrë nga zona e ndotur.
Për kontrollin e faktorëve, variablave,që ndikojne në 'kollë', nisur nga përfundimi I mësipër është vënë re se ka një ndikim të *zonës* dhe *konsumit* të duhanit në shtëpi, kur kalohet në zonën e ndotur. Nga studimet e më parshme[15], është vënë re dhe ndikimi I *'konsumit vetiak të duhanit*'. Kështu, që tek Kapitulli 5 , *Zbatim*,është paraqitur hollësisht, zbatimi i regresit logjistik, kur në modelim përdoret faktori *'zone, 'konsum_duhani' dhe 'konsum _duhani_ne_shtepi'.*

Në bazën tonë të të dhënave, variabli *kolle* është variabel kategorik. Ai pasi i është bërë një transformim rekordesh, është i koduar, 0 nëse personi nuk ka kolle dhe 1 kur ka kolle. Kështu ai është bërë nje variabël binar, që është modeluar me ane të modelit logjistik, për të kontrolluar ndikim apo jo të variablave të lartpërmendur.
Variabel i varur është konsideruar variabli *'kollë'*, variabël binar; ndërsa variablat me të cilën është ndërtuar parashikimi është variabli *'zone',area, 'konsum_duhan_në_shtëpi'* dhe variabli *'konsum_duhan'* , i trajtuar më parë si ndikues në *'kollë'* [14].
Modeli I komentuar gjërësisht tek Kapitulli 5 , nga kontrolli statistik ka dalë I vlefshëm. Janë komentuar koeficientët e modelimit, të paraqitura edhe tek Tabela 6-19. Por, duhet patur parasysh që me këto koeficientë, mund të modelohet lehtësisht variabli, *parashikim_kolle'.*

Tabela 6–19 Koeficientët e variablave në ekuacionin e modelimit

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	area	1.028	.326	9.967	1	.002	2.796	1.477	5.293
	pini_duhan	.147	.489	.090	1	.764	1.158	.444	3.018
	duhan shtepi	.122	.037	10.602	1	.001	1.130	1.050	1.216
	Constant	-2.267	.596	14.491	1	.000	.104		

a. Variable(s) entered on step 1: area, pini_duhan, duhan_shtepi.

Në output-et e afishuara merren edhe informacione se c'do të arrijë modelimi I fituar. Ne tabelën e kontigjencës sipas testit Hosmer and Lemeshow, tregohet se behet automatikisht nje ndarje e te dhenave ne gjashte grupe, e për secilin grup tregohet numri konkret i vlerave te vrojtuara të grupit dhe, numri i vlerave të grupit që siguron modeli. Krahasimi i këtyre vlerave tregon se per vlera të pritshme të ulta modeli nuk është aq i mirë.

Tabela 6–20 Tabela e klasifikimit pas modelimit

Classification Table[a]

Observed		Predicted		
		Kolle		Percentage
		jo	Po	Correct
kolle	jo	58	54	51.8
	po	37	70	65.4
Overall Percentage				58.4

a. The cut value is .500

Tabela e klasifikimit tregon se modelimi percakton saktesisht 51.8% te atyre që s'kane kolle, por 65.40% te atyre që kane kolle, ndërsa më parë, para modelimit, nuk arrihej të parashikohej asnje nga keto persona. Keshtu pergjindja totale e parashikimit me atë përfundim që jipej pa modelim është rritur me 18%

6.3 Analiza statistikore e problemeve hematologjike të banorëve të zonës (Problemi 3)

Ajri rrotull kësaj fushe, ku nafta është nxjerrë e përpunuar që me 1928 është i ndotur. Niveli i SO_2 është statistikisht i lartë (33 part per billion). Të dhënat e analizave të gjakut, përmbajnë rezultatet e analizave në gjak të banorëve të zonave në fjalë. Në këto rezultate vecohen eritrocitet, leukocitet dhe hemoglobina. Por, aty permbahen edhe të dhëna për parametrat e tjerë limfocite, monocite, granulocite, që përfshijnë neutrofilet dhe eosinofilet, që përbëjne formulen e leukocideve.
Për secilin person jane marrë të dhëna të përgjithshme si mosha, gjinia, sa kohë kanë që banojnë në zonë, historiku i ndonjë sëmundje në gjak. Database është ndërtuar nga Instituti i Shëndetësisë dhe i Shëndetit Publik, duke patur parasysh listën e tërë faktorëve që mund të shkaktojnë probleme hematologjike.

Rrezultatet e të dhënave në gjak për banorët e zonës së ndotur, u krahasuan me rezultatet e të dhënave koresponduese të zonës së kontrollit. Analiza statistikore, e sqaruar hollësisht në seksionin 1.4, synonte të kontrollojë nëse këto mesatare janë statistikisht të njëjta, apo jo. Gjatë kryerjes së procedurave te testit Leven, e më pas t_ testit, Studenti apo Welch, në varësi të rezultateve të testit të parë, afishohen dhe vlerat mesatare te parametrave në fjalë Tabela 6-21.

Tabela 6–21 Rezultatet e analizës deskriptive

Group Statistics

	area	N	Mean	Std. Deviation	Std. Error Mean
eritrocide	1.00	56	4718750.00	511578.884	68362.604
	2.00	84	4575588.24	1034865.505	177477.970
leukocide	1.00	56	7866.07	1768.130	236.276
	2.00	84	12029.41	29198.054	5007.425
hemoglobine	1.00	55	13.33	1.480	.200
	2.00	83	12.07	2.535	.441

Si rrezultat fuqisë së madhe kompjuterike të përpunimit të të dhënave, rezultatet e testeve shoqërohen edhe me afishimin e intervaleve të besimit të tyre, Tabela 6-22. Kështu, duke njohur edhe kufijtë normal ku duhet të lëvizin këto vlera, mund të lindin dyshime nëse këto kufij lëvizin jashtë normalitetit.

Tabela 6–22 Rezultatet e krahasimit të paramerave kryesor në gjak

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
eritrocide	Equal var. assumed	3.115	.081	.876	88	.383	143161.765	163447.949	-181656.688	467980.217
	Equal var. not assum.			.753	42.952	.456	143161.765	190189.052	-240403.509	526727.039
leukocide	Equal var. assumed	5.126	.026	-1.068	88	.289	-4163.340	3899.244	-11912.269	3585.588
	Equal var. not assum.			-.831	33.147	.412	-4163.340	5012.996	-14360.640	6033.959
hemoglobine	Equal var. assumed	.625	.431	2.931	86	.004	1.253	.427	.403	2.102

	Equal var. not assum.			2.586	45.301	.013	1.253	.484	.277	2.228

Nga analiza e tri parametrave eritrocite, leukocite dhe hemoglobinës u morrën këto afishime:

- Tabela e mësipërme tregon se, për eritrocitet plotesohet kushti i homogjenitetit te variances së popullimit, dhe me pas, vërehet se s'ka dallim statistikor midis mesatares eritocideve të grupit te ndotur dhe atij të kontrollit.
- p- vlera .412 tregon se niveli i leukocideve është i njëjtë në dy grupet në studim, nga ana statistikore, por nuk plotesohet kushti i homogjenitetit të variances.
- p-vlera prane hemoglobinës .004 , dhe vlerat e afishuara te mesatares së diferencës, tregojnë se kjo vlerë është më e ulët në zonën e kontrollit.

Niveli mesatar i eritrocideve në të dy grupet e analizuara duket se është nën kufijtë normal. Duke patur parasysh dhe problemet për të cilat zona përmendej, u shtrua problemi ka apo jo shenja të anemisë në gjak. Sipas specialistëve kufiri përcaktues për të ishte 3500000. Pra u shtrua hipoteza:

hipoteza zero (ka shenja të anemisë)

$H_0 : \mu=3500000$

hipoteza alternative (nuk ka shenja të anemisë)

$Ha: \mu > 3500000$

Testi, outpu-et mund të lexohen në Tabelën 6-23, vlera e statistikës t që jep një p-vlerë 8.427E-30, tregojnë se nuk ka shenja të anemisë.

Tabela 6–23 Kontrolli hipotezës për kufirin minimal

One-Sample Test

	Test Value = 3500000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
eritrocide	15.886	136	.000	1128971.963	988070.29	1269873.64

Por, testi i hipotezës tjetër me vlerën 4500000 që sipas specialistëve është kufiri më i ulët i lejuar i këtij parametri

hipoteza zero $H_0 : \mu=4500000$
hipoteza alternative $Ha: \mu >4500000$

tregon se niveli I eritrocideve është në kufijtë minimum, të vlerave normale me nivel besimi 0.1 (P=0.07) , Tabela 6-24.

Tabela 6–24 Kontrolli hipotezes per minimumin e vleras normale

One-Sample Test

	Test Value = 4500000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Loëer	Upper
eritrocide	1.815	136	.072	128971.963	-11929.71	269873.64

Sic u përmend edhe më siper , në vlerat e komponentit të tretë në gjak është vënë re një diferencë e vogël , po signifikative midis dy zonave në shqyrtim. Vlera e t-testit dhe p-vlera tregojnë influencën e zonës, në këte komponente të gjakut. Nisur nga përvoja e specialistëve, duke shfrytëzuar të dhënat e përgjithshme, u kontrollua efekti i dy faktoreve 'gjini' dhe 'area'. Rezultatet e paraqitura në Tabelën 6-25 tregojne ndikimin e dy faktoreve ne shqyrtim. p-vlerat 0.000012 pranë *'zone'*-s, dhe tej mase më e ulët, pranë *'gjinise* ' tregojne per efektin e dy faktoreve në fjale përqindja e femrave në të dhenat Data Analizagjak. U vure re se përqindja e femrave në zonën industriale ishte (61.2%), ndërsa vlera e kësaj përqindje në zonën e kontrollit ishte (48%).

Tabela 6–25 Anova dypërmasore

Tests of Between-Subjects Effects

Dependent Variable: hemoglobine

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Intercept	68859.115	1	68859.115	32842.165	.000
area	41.261	1	41.261	19.679	.000
gjinia	115.873	1	115.873	55.265	.000
area * gjinia	.354	1	.354	.169	.681
Error	853.344	407	2.097		
Total	70295.400	411			
Corrected Total	1007.894	410			

a. R Squared = .153 (Adjusted R Squared = .147)

Një analizë statistikore e detajuar u bë për komponentët e tjerë të gjakut, limfocite, monocite dhe granulocite (neutrofile dhe esonofile).

Studimi i tabelës së rezultateve, Tabela 6-26, tregon se, mesataret e gjithë parametrave të formulës së leukocideve janë me të ulta në zonën industriale se ato të nivelit mesatar .

Tabela 6–26 Analiza deskriptive

Groups	Formula e leukocideve	Statistika deskriptive			
		Mesatarja matematike	Devijimi standart	Q1	Q3
Zona kontrollit	Limfocide	33.46	1.2000	25.75	41.00
	Monocide	5.920	0.226	5.0000	7.0000
	Granulocise neutrofile eosinofile	 59.03 2.9800	 1.56 0.147	 53.95 2.00	 64.11 4.00
Zona industriale	Limfocide	24.04	2.60	0.37	37.75
	Monocide	4.149	3.3.7	0.070	7.000
	Granulocise neutrofile eosinofile	 57.04 2.1150	 1.36 0.2490	 50.50 0.5200	 64.75 3.000

Rezultatet e metodës Anova një-përmasore, Tabela 6-27, për *limfocitet, neutrofilet* dhe *eosinofilet* (P= 0.002, F=9.98) , ndotjes, në këto komponente të gjakut.

Tabela 6–27 Anova njëpërmasore per komponentet e tjera të gjakut

Tests of Between-Subjects Effects

Dependent Variable: monocide

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	17.271	1	17.271	2.551	.114
area	346.804	2	173.402	25.617	.000
Error	568.608	84	6.769		
Total	1481.180	87			
Corrected Total	915.411	86			

a. R Squared = .379 (Adjusted R Squared = .364)

Tests of Between-Subjects Effects

Dependent Variable: limfocide

Source	Type III Sum of Squares	df	Mean Square	F	Sig.

Intercept	583.342	1	583.342	2.433	.122
area	11900.009	2	5950.005	24.821	.000
Error	20855.661	87	239.720		
Total	53230.089	90			
Corrected Total	32755.670	89			

a. R Squared = .363 (Adjusted R Squared = .349)

Faktoret që mund të ndikojnë në shfaqjen e problemeve hematologjike për banorët, janë listuar nga Instituti i Shendetit Publik dhe në bazë tyre është hartuar dhe nje pyetsor, me informacione të detajuara për *'moshë'*n*,'histori te tilla sëmundjesh në familje', ' koheqëndrimi në* zonë'.
Qëllimi është analiza dhe modelimi i nje modeli statistiko-probabilitar që të shqyrtojë efektin e tyre në cdo komponente gjaku, duke synuar që këto parametra të modelojnë komponentin e shqyrtuar.

Për këtë u kthyen variablat në shqyrtim 'mosha' dhe koheqëndrimi në zonë ' në variabla kategorik, me vlera të përcaktuara si më poshtë:

$$kohëqëndrim\_kategorik=\begin{cases}0, & nqs\ personi\ \ ka\ më\ \ pak\ se\ 20\ vjet\ banorë\\ 1, & nqs\ personi\ \ është\ 20-40\ vjet\ banorë\\ 2, & nqs\ personi\ \ ka\ me\ shume\ se\ 40\ \ vjet\ banore\end{cases}$$

dhe

$$mosha\_binar=\begin{cases}0, & nqs\ person\ \ nën\ 25\ vjeç\\ 1, & nqs\ person\ mbi\ 25\ vjeç\end{cases}$$

Meqë komponentet e formules e leukocideve janë të gjithe n. r. të vazhdueshme, dhe kërkohet të shqyrtohet efekti i n.r të tipit kategorik, një metodë e ndjekur vazhdimisht në studimin e problemeve të tilla përfshin kthimin e vlerave të parametrave të formulës lekoucidare në n.r kategorike. Pra nga parametrat e formulës së leukocideve, janë krijuar kater variabla binar yi, për $1\leq i\leq 4$. Ato përcaktohen 1 nqs vlera e parametrit perkatës është me e vogël se kuartili i pare, dhe 0 në rastet e tjera. Është përdorur kuartili i pare si vlere përcaktuese e funksioneve në fjalë. Kështu testohet nëse kapja e këtyre vlerave nën kuartil, ndikohet nga faktorët në shqyrtim, e për me tepër cilët faktorë janë ato që rrisin rezikun , riskun, e vlerave te tilla, jo të mira.
Me anë të regresit logjistik është shqyrtuar ndikimi i 'mosha_ binar' dhe 'koheqëndrimi _kategorik' në,' zonë' në komponentët e formulës së leukocideve dhe rezultatet janë paraqitur në Tabelën 6-28.

Tabela 6–28 Regresi logjistik për indikatoret rasteve problematike

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	mosha_binar	-.124	.125	5.264	1	.035.	.674
	kateg_qendrim	-.642	.271	5.598	2	018	526
	Constant	.638	.636	1.008	1	.315	1.893

a. Variable(s) entered on step 1: kategqendrimi, mosha_binar

Tabela 6-28 tregon për ndikim te dy variablave (p-vlerat janë.025. dhe .018) mosha_binar dhe kateg_qendrim . Koeficientat negativ do të thotë se ulja e parametrit mosha_binar, pra kur flitet për banorë nën 25 vjec , do të rrise logit e probabiliteti, pra mundësitë e shfaqjeve të parametrave të ulta janë më të mëdhja se mundëstë e shfaqjes së vlerave normale.

Po kështu, mund të përqëndrohemi edhe në faktorin tjetër të rëndësishëm të këtij modelimi, ulja e kategorisë së qëndrimit, pra banoret që s'janë rezidentë për një kohë të gjate, kemi një rritje të mundësive per kapjen e vlerave të komponentëve të gjakut nën normalen.

Pra, në moshat nen 25 vjec ose tek banorët me nje kohë jo shumë të madhe qëndrimi mundësitë, që shënjat e parametrave të ulta të formulës leukucidare të shfaqen, janë më të mëdha.

Gjithë parametrat e formulës së lekucideve janë variabla të pavarur të vazhdueshëm. Mosha_binar (dy nivele) and kohëqëndrimi (tre nivele), mund të shihen si faktore. Kështu mund të zbatohet edhe procedura e Anova-Dypërmasore për të kontrolluar se *kohe qendrimi* dhe *mosha_binar* kanë ndikim në parametrat e shqyrtuara, ose jo dhe të kontrollohet ndërthurja e faktoreve.

Kontrolli i vlerave të afishuara (Tabela 6-29) tregon :

Tek *monocidet*: Kemi ndikim të *kohëqendrimit* p_value 0.010, por jo ndikim të *'mosha-binar'* Vlera e statistikes F=.257 dhe p_value .774

Vërehet se efekti nderthurës i dy faktorëve në shqyrtim është josignifikativ (ky efekt paraqitet më qarte ne Tabelen 29, p_value 0.324, në afishimeve per monocidet)

Tek *ezinofailet* ndikimi I kohe_qëndrimit është i pranueshëm me nivel besimi 0.1. P_value e testit ANOVA për këtë variabël është .073 . Dhe, gjithashtu I qarte eshte dhe ndikimi i rolit nderthurës i dy faktorëve në shqyrtim (p_vlera .018).

Tek *limfocidet* ndikimi i kohë_qëndrimit, shihet se pas kontrollit të metodës ANOVA . Ne tabelen 6-29 p_vlera e afishuar është 0.004. Por, ndryshe nga parametrat e tjerë të formulës së leukocideve këtu, pra për këtë bazë të dhënash, nuk vërehet ndërthurja mes dy faktorëve në shqyrtim.

Kështu, përfundimisht, ekzistojnë te paktën dy grupe kohë_qëndrimi ku verehet ndryshimi I komponentëve të formules së leukocideve. Për përcaktimin e grupit të '*kohëqëndrimi*t' me ndryshim të dallueshem të komponentit në gjak është përdorur metoda Tukey (variabli kategorik kohëqëndrim' është me tri kategori ndaj kjo metodë jep vlera më të mira).

Përdorimi I post hoc Tukey, për komponentin *monocide* , tregon se deri në 20 vjetët e para të koheqëndrimit niveli I uljes se monociteve eshte signifikativ(p=0.04), por me pas ai përmirësohet dhe është jo signifikativ pas 40 vjet kohëqëndrimi (p= 0.21).

Sasia e *limfociteve* të banorëve për tri nivelet e kohëqëndrimit është respektivisht: 24.3 (për banorët që janë 1-20 vjet në këtë zonë); 23.1 (për banorët që janë 21-40 vjet në këtë zonë); 22.3 (për banorët që janë mbi 40 vjet në këtë zonë).

Për përcaktimin e grupeve me vlerë mesatare të ndryshme, te parametrave të gjakut, të përdoresh Post Hoc e ANOVA së komponentes së gjakut ndaj *kateg_kohëqëndrimi*, do të thotë ta ndertosh modelin pa marrë parasysh rolin ndërthurës së *moshës_binar* me *kateg_kohëqëndrim*.

Modelimi do të ishte më korrekt, për *ezinofailet* nëse kjo ndërthurje do të merret parasysh. Ndaj zgjidhjen e marrim duke ndërhyrë në sintaksën e SPSS, e cila mundëson ekzekutimin e komandave që analizojne modelin, duke patur parasysh këtë faktor të ndërthurjes. Sintaksa e shkruar në SPSS është bërë si më poshtë:

```
DATASET ACTIVATE DataStudimi I analizes se gjakut zona industriale.
UNIANOVA ezinofile BY mosha_binar kategqendrimi
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /POSTHOC=mosha_binar kategqendrimi(SCHEFFE)
  /PLOT=PROFILE(kategqendrimi*mosha_binar)
  /EMMEANS=TABLES(mosha_binar) COMPARE ADJ(LSD)
  /EMMEANS=TABLES(kategqendrimi) COMPARE ADJ(LSD)
  /EMMEANS=TABLES(mosha_binar*kategqendrimi)COMPARE ADJ(mosha_binar)
  /PRINT=ETASQ HOMOGENEITY DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /DESIGN=mosha_binar kategqendrimi mosha_binar*kategqendrimi.
```

Rezultatet e *ezinofaileve*, Tabela 6-30, tregojnë :

- Ne vitet e para të kohëqëndrimit vlera e limfocideve mesatare është përkatësisht .003 dhe .0035 për dy nivelet e moshave, pra është nen kufijtë minimal që është më pak se 0.35 %.
- Por tabela 6-30, e afishuar nga ndërhyrja në sintakse, duke shfaqur një diferencë negative, tregon se efekti i ndryshimit në limfocide vërehet shumë

mes grup moshës së re e tjetrës, në nivelin e parë të kohëqëndrimit. Vlera negative mesatare tregon madje, se vlera mesatare është me e ulet tek te rinjtë.

- Kjo diferencë mes grupmoshave nuk është më signifikative në nivelet e tjera të kohëqëndrimit.

Fillimisht mund të konkludohet, rendesia e përdorimit të efektit ndërthurës mes dy vektoreve në shqyrtim.

Përfundimisht,nga krahasimi I përfundimeve të regresit logjistik me variablin parametrat e gjakut të kategorizuar, dhe modelit te fundit të analizës, mund të themi se kthimi i një variabli të vazhdueshem në një variabël kategorik shoqërohet me tre probleme;

- *Së pari*, problemi kontrollit të testeve të shumëfishta, e kthen në krahasim të cifteve të kuartileve.

- *Së dyti,* përdorimi i metodes kërkon plotësimin e kushtit të homogjenitetit brenda grupit të krijuar, duke zëvendësuar kushtin e homogjenitetit mes grupeve të variablit origjinal të vazhdueshëm. Statistikisht ky kusht, për ndryshoren e rastit të vazhdueshme, është i njëjtë me kushtin e homogjenitetit të të gjithë popullimit.

- *Së fundmi,* vështirësi paraqet punimi me këtë metode, sepse pengon shumë krahasimin e rezultateve te punimeve me njëri tjetrin, sepse këto pika përcaktuese, prerëse që përdoren për të karakterizuar ndryshoret e rastit të reja mund të jenë të ndryshme në studime të ndryshme.

Tabela 6–29 Analiza dy përmasore

Tests of Between-Subjects Effects

Dependent Variable: monocide

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.004[a]	6	.001	2.416	.072
Intercept	.020	1	.020	72.340	.000
mosha-binar	.000	2	7.131E-5	.257	.774
kategqendrimi	.004	3	.001	4.306	.010
mosha-binar * kategqendrimi	.000	1	.000	.997	.324
Error	.012	43	.000		
Total	.185	50			
Corrected Total	.016	49			

a. R Squared = .252 (Adjusted R Squared = .148)

Dependent Variable: limfocide

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.133[a]	5	.027	3.553	.009
Intercept	.794	1	.794	106.105	.000
mosha_binar	.003	1	.003	.390	.536
kategqendrimi	.113	3	.038	5.041	.004
mosha_binar * kategqendrimi	.005	1	.005	.723	.400
Error	.329	44	.007		
Total	5.874	50			
Corrected Total	.462	49			

a. R Squared = .288 (Adjusted R Squared = .207)

Dependent Variable: ezinofile

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.002[a]	6	.000	2.512	.086
Intercept	.004	1	.004	38.640	.000
mosha-binar	.000	2	6.502E-5	.621	.542
kategqendrimi	.001	3	.000	3.714	.018
mosha-binar * kategqendrimi	.000	1	.000	2.386	.073
Error	.005	43	.000		
Total	.049	50			
Corrected Total	.006	49			

Tabela 6–30 Kontrolli i efektit nderthurës tek ezinofailet

Pairwise Comparisons

Dependent Variable: ezinofile

kategqendrimi	(I) mosha_binar	(J) mosha_binar	Mean Difference (I-J)	Std. Error	Sig.[c]	95% Confidence Interval for Difference[c]	
						Lower Bound	Upper Bound
0-20	nen25	mbi 25	-.006[a]	.004	.020	-.004	-.002
	mbi 25	nen25	.006[b]	.004	.042	.002	.004
20-40	nen25	mbi 25	.004	.005	.457	-.007	.015
	mbi 25	nen25	-.004	.005	.457	-.015	.007
mbi 40	nen25	mbi 25	-.003	.004	.494	-.011	.005
	mbi 25	nen25	.003	.004	.494	-.005	.011

Based on estimated marginal means

6.4 Analiza statistikore e ndikimit te zones ne bujqesi

6.4.1 Baza e të dhënave

Fushat e Patos Marinëz dhe Kucovës janë zonat më te rëndësishmë industriale në vend, pra zona ku perqëndrimi i gazeve ndotës,si SO_2, është I lartë. Por, këto zona shfrytëzohen edhe si zona bujqësore.Nga projekti ANFI I Ministrisë së Mjedisit, i financuar nga Banka Botërore, u morën informacione për mbulesën bimore ne pergjithesi dhe shtrirjen e klasifikin e pyjeve te vendit. Mbi bazë të këtyre informacioneve në ne programin ArcGIS 10 u realizuan hartat e mbulimit bujqësor të zonave.

Të dhënat janë marrë nga tre fshatra. Dy prej tyre Ballsh-Marinëz dhe Sheqishtë-Belinë janë në zonën industriale, ndërsa tjetri në zonën joindustriale të Kurjanit.Sic thamë tek shtrimi i problemit, Figura 6-2 të dy fshatrat e parë mund të konsiderohen njelloj të ekspozuar nga SO_2 , ndërsa i treti si zonë e paster, sepse nuk ka perqendrim të puseve ne të. Kështu në studim Kurjani është konsideruar si zona e *kontrollit,* ndërsa Sheqishtë- Belinë dhe Marinëz është konsideruar si zonë e *ndotur.*

Në zonat e mbjella me *bimëve bujqësore* të këtyre fshatrave (Figura 6-4), janë marrë informacione për prodhueshmërinë e shtatë kulturave bujqësore, nga Kurjani (në zonën e *kontrollit*) dhe pjesa tjetër nga Sheqishtë- Belinë dhe Marinëz (në zonën e *ndotur*). Informacionet e marra lidhen me sasinë e prodhimit të grurit, misrit patates, perimeve, fasules, frutave dhe jonxhës. Është vëzhguar dhe distanca e shtëpive nga puset e naftes në zonë, tipi i sistemit të ujrave të zeza të përdorura, menyra e grumbullimit të mbeturinave dhe niveli i përkujdesjes së tyre për produktet bujqësore.

6.4.2 Analiza statistikore e problemit

Të dhënat per prodhueshmërinë e shtatë produkteve bujqësore janë analizuar dhe, nga analiza deskriptive e të dhënave të zgjedhjes janë marrë rezultatet e paraqitura ne *tabelen 6-31*

Figura 6–4 Harta e bimësisë në zonën në studim

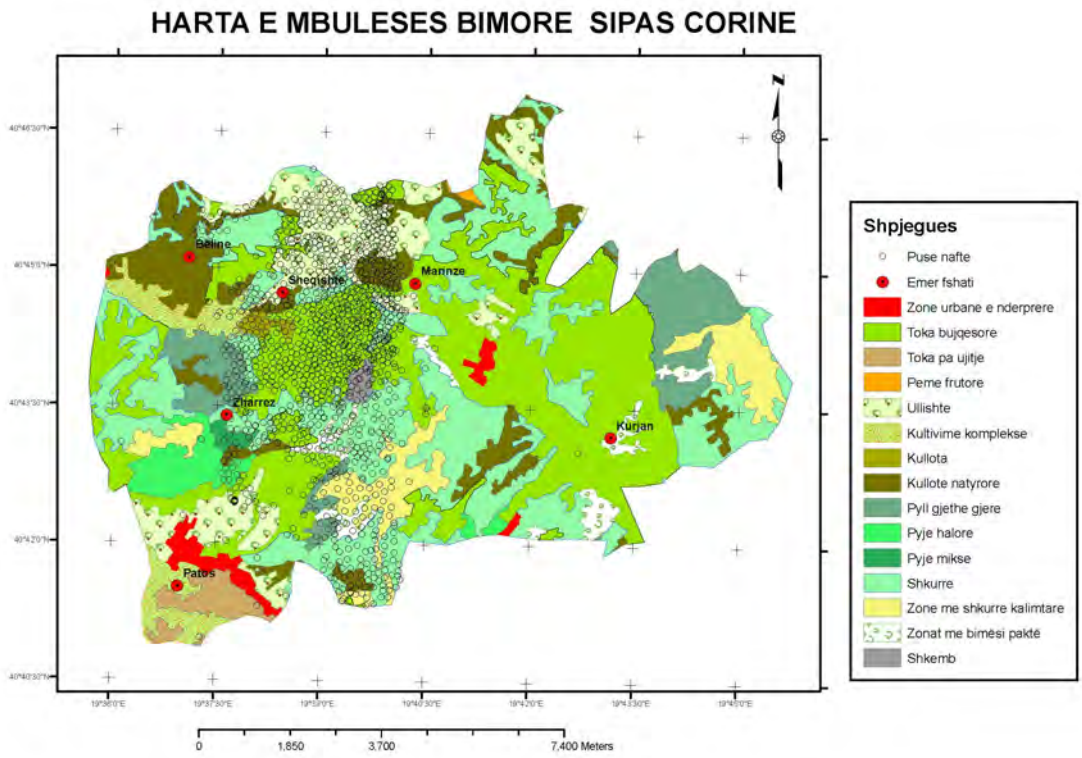


Figura 6–5 Harta e rrjetit hidrologjik

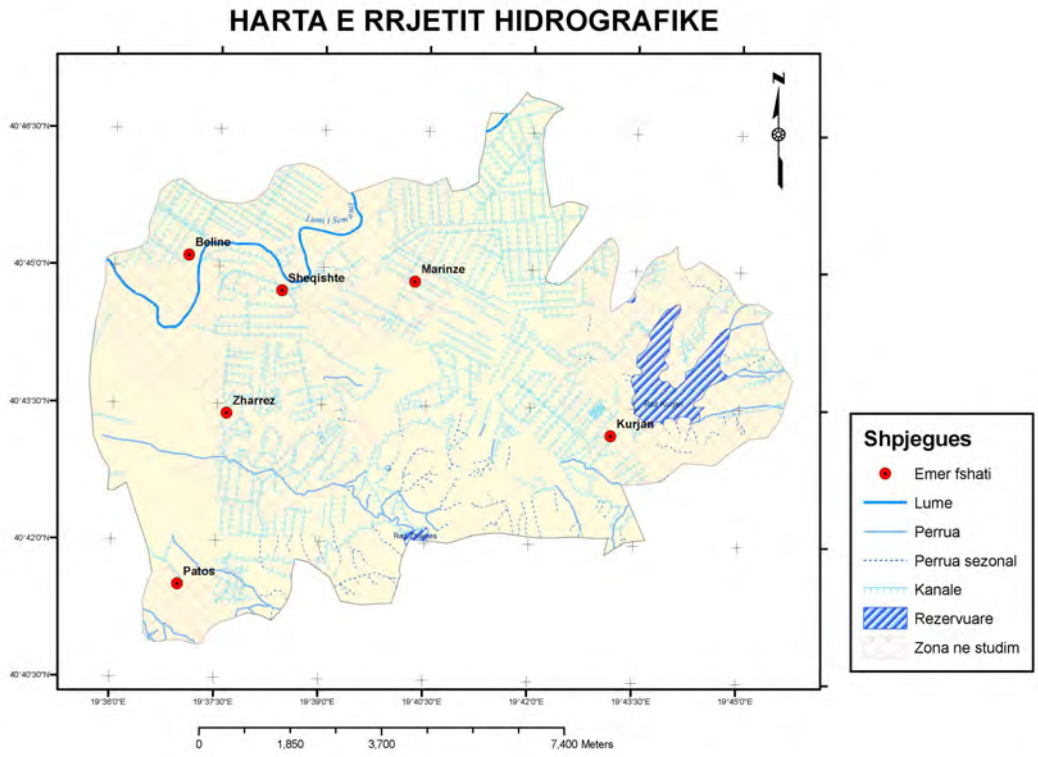


Tabela 6–31 Analiza deskriptive e bimëve

Descriptive Statistics

Zona		N	Mean	Std. Deviation
Zona e kontrollit	grure	45	829.3333	331.24009
	miser	45	651.1111	179.16755
	patate	45	155.3333	156.05360
	perime	45	170.0000	108.39742
	jonxhe	45	2104.4444	1723.23105
	fasule	45	67.7778	37.15073
	fruta	45	184.4444	158.04999
Zona e ndotur	grure	110	812.9091	545.69263
	miser	110	630.9091	225.34192
	patate	110	67.2727	39.35044
	perime	110	110.6364	51.79843
	jonxhe	110	4510.0000	6030.83742
	fasule	110	42.5455	24.36106
	fruta	110	157.0000	242.22800

Duke pare tabelen 6-31, shihet se:

- *Mesatarja e prodhimit* te misrit, patates, perimes dhe fasules është më e ulët në zonën industriale.
- *Mesatarja e prodhimit* të frutave dhe jonxhës duket të jetë më e lartë në zonën e ndotur.

Kontrolli i kushtit të homogjenitetit të variances, testi Levin, nuk jep te njejtin pergjigje per të gjithe komponentet e vektorit shtatë përmasor *Bimësi*, ndaj krahasimi I mesatareve te komponenteve te vektorit *Bimë* nuk është bërë me metoden ANOVA, por me anë të t- testit, sepse, sic dihet, kjo metodë bën të mundur vlerësimin e dy mesatarëve pavarësisht plotësimit apo jo te kushtit në fjalë.

Në tabelen 6-32 janë paraqitur vlerat e *t*-testit Studenti apo *t*-testit Welch, në varësi të faktit qe komponenti I shqyrtuar e ploteson apo jo kushtin e homogjenitetit të variancës. Vlerat e statistikës F te testit Levin dhe p-vlerat koresponduese, tregojnë se varianca në zonat e ndotura është më e madhe për misrin(p-vlerë< 0.001), pataten (p-vlerë < 0.001), perimet (p- vlerë < 0.001), fasulen (p- vlerë < 0.001) dhe jonxhen (p-vlerë 0.028). Ky rezultat e ka dhe një farë justifikimi statistikor, sepse vellimi I zgjedhjes ne zonën e ndotur eshte me i madh.

Tabela 6–32 Rezultatet e testit Levin dhe te t-testit për mesataret

	Levene's Test for Equality of Variances		**t-test for Equality of Means**				
Bimë	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
grure	.234	.629	.188	153	.851	16.424	87.35

miser	2.447	.120	.536	153	.593	20.202	37.70
patate	31.638	.000	3.737	46.306	.001	88.061	23.56
perime	21.773	.000	3.513	52.420	.001	59.364	16.90
jonxhe	4.941	.028	-3.820	142.163	.000	-2406	629.79
fasule	16.393	.000	4.202	60.051	.000	25.232	6.01
fruta	1.313	.254	.701	153	.485	27.444	39.16

Përdorimi *Independent-Simple T Test* në SPSS bën të mundur që, duke shfrytëzuar rezultatet e kesaj metode në varësi te përgjigjeve të testit Levin, të paraqitura në anën e djathtë te tabelës [32], të arrihet në këto përfundime:

- Ka ndryshim statistikor mes mesatareve te prodhimeve tek *patatja, fasulja* dhe *perimet*.
- Nuk ka ndryshim statistikor tek *frutat* (p-value 0 .701), *gruri* (p-value 0.188) dhe *misri* (p-value 0.536).
- Mesatarja e prodhimit te *jonxhes* është statistikisht më e lartë ne zonën e ndotur, derisa diferenca e studiuar e vleres se komponentit në fjalë në zonën e ndotur nga zona e kontrollit është negative.
- Mesatarja e *patates*, *perimes* dhe *fasules* është statistikisht më e ulët në zonën e ndotur.

Tabela 6–33 Rezultatet e analizës së kontrastit

Contrast Coefficients

	area	
Contrast	controlled area	polluted area
1	1	-1.3

Contrast Tests

Bimë		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Patate	Does not assume equal variances	1	67.8788[a]	23.768	2.856	47.916	.006
Perime	Does not assume equal variances	1	26.1727[a]	17.387	1.505	58.402	.038
Fasule	Does not assume equal variances	1	12.4687[a]	6.307	1.977	71.498	.052

a. The sum of the contrast coefficients is not zero.

Më pas me anë të analizës së kontrastit mes dy grupeve për secilin nga tre komponentet që kanë prodhueshmëri më të ulët , pasi u perdorën koeficientët 1 dhe -1.3 , u provua se per të tre bimët prodhueshmëria në zonën e kontrollit është statistikisht sa 1.3 e prodhueshmerise në zonën e ndotur..

Programi SPSS mundëson llogaritjen e fuqisë së testit duke përdorur opsionin *the observed power of t-test*, me anë të efektit specifik te vëllimit te zgjedhjes, vëllimit të zgjedhjes, dhe p-vleres (significance).Duke analizuar këtë fuqi për tërë bimët arrihet në konkluzionin:

- Fuqia e testit për *perimet* është 0.96, për *fasulen* , *patate*n 0.99 dhe për *jonxhën* 0.788.
- Fuqia e vrojtuar e testit për *frutat* t-test është 0.11; e *grurit* t-test 0.053, dhe e *misrit* t-test është 0.083.

Kështu mund të konkludohet se po të kishim një vëllim zgjedhje më të madh , rezultati mund të ishte më signifikativ, duke patur parasysh ndikimin pozitiv te rritjes së vellimit të zgjedhjes ne fuqinë e testit. Për *grurin*, që sic lexohet nga tabela 6-32 e të dhenave deskriptive, ka devijim standart të madh, kjo mund të ndikonte në ngushtimin e intervaleve të besimit [33] , duke cuar ndoshta përfundimisht në një vlerësim më të mirë të vlerave të popullimit.

Por duhet theksuar se fuqitë e ulta të testit ne rastet e *grurit, misrit* dhe *frutave*, nuk do të thone se konkluzioni ynë,që nuk kishte ndryshim prodhushmeria e ketyre produkteve ne zonat e studiuara, nuk është i vertetë.

Problemi i fuqise së testit ia vlen me shumë të studiohet nese pergjigja e *t*-testit ka qenë signifikative, pra në rastin tonë ka treguar për ndikim të zonës në komponentin e studiuar bimë.Kështu, rezultatet e afishuara tregojnë një test më të dobët për *jonxhën*, sesa per *patate*n, *perimen* dhe *fasulen.*

Sasia e SO_2 që ndryshon ne këto dy tipe zonash dhe influencon ne produktivitetin e bimëve është faktor abiotik. Faktorët abiotik janë ato faktorë, fizik dhe kimik jojetësor, të cilët ndikojnë tek bimët. Por ka dhe një sërë faktorë të tjerë të tillë si zona ku luhatet temperatura, lagështia e tokës, niveli I ndotjes, të cilët ndikojnë në këtë prodhueshmëri.

Temperatura është konsideruar e njëjtë në të dy sipërfaqet sepse zonat në fjalë janë të lokalizuara jo larg njëra tjëtrës.

Duke shfrytëzuar Bazën e të Dhënave të Rrjetit Hidrologjik Kombetar, të siguruar nga një projekt tjeter, Watershed, qe ka studiuar basenet lumore dhe ka siguruar dixhitalizimin e të gjitha hartave topologjike të Shqipërisë, në GIS, u bë harta e rrjetit hidrologjik të zonave, Figura 6-5. Shqyrtimi i saj tregoi, se sistemi i kanalizimeve është i pasur në zonat në shqyrtim dhe nuk ka ndonje dallim mes dy zonave përsa i perket problemit të lagështisë së tokës.

Niveli i ndotjes, në pamundesitë e matjes me aparutura speciale të kontrollit të zonave, mendohet të varet nga distanca e shtëpive të intervistuarve nga puset e naftës. Vlera e kësj distance në metra përcaktojnë variablin numerik, *distanca*.

Për të studiuar edhe variabla të tjere të cilet ndikojnë ne problemin e prodhueshmërisë se bimëve, janë përcaktuar edhe dy variabla të tjerë:

ujrat e zeza dhe *mbetjet urbane* u studiuan si variabla kategorik të koduara:

$$\textit{ujrat e zeza}=\begin{cases}1 & \text{në } rast\ kanalizimesh\\ 2 & \text{në } rastet\ e\ \ gropat\ septike\\ 3 & \text{në } rastet\ e\ përzierjes\ me\ ujin\ e\ rrjedhshëm\end{cases}$$

$$\textit{mbetjet urbane}=\begin{cases}1 & \text{në } rast\ transporti\ të\ mbetjeve\\ 2 & \text{në } rast\ groposjesh\ se\ tyre\ ne\ tokë\\ 3 & \text{në } rastin\ e\ djegjes\ së\ mbetjeve\end{cases}$$

Variablat *ujrat e zeza* dhe *mbetjet urbane* jane parë të lidhur me faktorët abiotik të tipit të dheut.
Të tri keto variabla janë trajtuar si variabla të pavarur. Të dhënat për to, si dhe të dhënat për prodhueshmerinë e bimëve janë analizuar me metoden e regresit per të identifikuar faktorë të tjerë abiotik që veprojnë tek bimët dhe mund të ndryshojnë ndikimin e SO_2 tek to.
Për të testuar nëse këto tri variabla ndikojnë në parashikimin e sasisë së prodhimit te *patates, fasules, perimeve* dhe *jonxhes* është pëdorur metoda e regresit hirarkik të shumëfishtë[34]. Si variabli i pavarur i listuar më parë, *the independent variable listed first*, u përdor variabli *zone*, dhe si variabla të shtuar , *the added variables ,distance, ujrat e zeza* dhe *mbetjet urbane*.
Ndërsa për të studiuar efektin tek *gruri, misri* dhe *frutat* u perdor metoda e zakonshme e regresit linear, sepse nga kontrolli i bere ketu nuk u vu re ndikim i variablit *zone*.
Probabilitetet e statistikes F per te gjitha regresionet e studiuara janë paraqitur në Tabelen 6-34, dhe tregojnë se ky model i regresit të shumëfishtë duhet marre parasysh tek *perimet , jonxha, gruri, misri* dhe *frutat*, derisa vlerat e statistikes së siperpermendur kane qene josignifikative. Kështu hipotezat zero që nuk kanë lidhje gjithë variablat e pavarur me variablin e varur, *prodhushmeria e bimes*, bie poshtë .
Ne tabele janë paraqitur dhe vlerat e statistikes *R2- change*. Informacioni i marrë nga variablat e shtuar, redukton parashikimin në prodhueshmerinë e bimëve (20% per *perimet*, 4.3% for për *jonxhën*, 7.6% për *grurin*, 5.4% per *misrin*, and 6.7% për *frutat*).

Tabela 6–34 Rezultatet e regresit hirarkik dhe regresit të shumëfishtë të zakonshëm

Bimë	**F-value**	**Sig.**	**R-Square**
Perime	2.601	.05	0.199
Jonxhe	2.3	.07	0.043
Fasule	.1.44	.233	
Patate	1.591	.194	
Grurë	3.092	.018	0.076
Misër	2.126	.08	0.054
Fruta	3.743	.06	0.067

Duke analizuar tabelen e koeficientëve të regresit të shumëfishtë, shihet efekti i variablit *mbetje urbabe*. Hipoteza zero b=0[35] , nga afishimet e marra hidhet poshtë , duke treguar se nuk ka ndikim ne komponentet e vektorit bimë. Por t-vlera 2.746 (sig 0.009) për *jonxhën*, tregon për një farë lidhje të mundshme. Koeficientët pozitiv 10.801 dhe 1375 të modelit të regresit të linear si dhe drejtimi i kodimit të variablit *mbetje urbane* , tregon për një farë ndikimi pozitiv të trajtimit sa më të mirë të mbetjeve në rritjen e prodhimtarisë.
Ndikimi i faktorit tjetër abiotik *ujrat e zeza* duhet marrë parasysh, sepse tek *gruri* (t-test -3.415 dhe sig. 0.01), *misri* (t-test -2.692 dhe sig. 0.08) dhe *frutat* (t-test -3.787 dhe sig.< 0.01). Vlerat negative të koeficientëve përkatësisht -436, -150.5 dhe -215 ne modelin e regresit linear përcaktojnë drejtimin e lidhjes. Duke verejtur kodin e

variablit *ujra të zeza*[36], ndërsa kodi rritet nuk ka asnje proces sanitar në trajtimin e ujrave të zeza, arrihet në përfundimin se një sistem më i mirë i ujrave të zëza ndimon në prodhueshmëri me të lartë tek disa bime.
Analiza e rolit të faktorëve abiotik tek bimët, tregoi se efekti i dioksidit të squfurit reduktohet si rezultat i ndikimit të faktorëvë të tjerë mbetje urbane dhe sistemi i ujrave të zeza.

6.5 Përfundime

Duke përmbledhur rezultatet në analizën e përfundimeve mund të themi:

- Testi Leven kontrollon barazimin e variancave në të gjitha nivelet e variablit të pavarur, të cilat përcaktohen nga variablat e pavarur. Afishimet e ketij testi (the spread-versus-level) përdoren për të kontrolluar testin dhe me pas duke patur parasysh kushtet perkatëse të secilës metode, aplikohet metoda e duhur në krahasimin e mesatareve.
- Llogaritja e fuqisë se testit na ndimon t'i tregojmë një kujdes më të madh të dhënave, por nuk tregon në rastet vecanerisht të një rezultati jo signifikativ, se vendimi i marre nuk është i drejtë. Ai ia vlen te studiohet në rastet kur përgjigja e testit është signifikative.
- Tipi i metodes së regresit të shumefishtë që duhet aplikuar, përcaktohet nga tipi i pyetjes te ciles ajo duhet ti përgjigjet.

Studimi i problemit të ndikimit te dioksidit te sulfurit tek bimët tregoi se:

- Kur substanca ne fjalë clirohet nga siperfaqe te mëdha si zona industriale, nga kontiniere, puse ose fuçi, sasia e prodhimit tek disa bimë reduktohet mesatarisht 13%.
- Bimët sillen në mënyrë të ndryshme me dioksidin e squfurit, meqë kanë një ndyshime përsa I përket mënyrës që ato e thithin gazin dhe aftësi të ndryshme në detoksifikimin e ndotësve dhe largimin e sasive të tepërta të squfurit. Konkretisht vetem tek patatja, fasulja dhe tek perimet sasia e gazit të cliruar në këto zona ndikonte negativisht.
- Faktoret abiotik sistemi i ujrave të zeza dhe, ne menyre jo shumë te sigurt dhe ai i mbetjeve urbabe, kanë ndikim ne produktivitetin e bimeve. Sa më shumë kujdes të tregohet në to nga ana sanitare, aq më pozitivisht ato ndikojne tek bimët.

7 Shtojcë - Analiza statistikore e të dhënave për përcaktimin e faktorëve që ndikojnë në memorje

7.1 Disa fjalë për studimin

Këto ditë, gjithmonë e më shumë njerëz po zgjedhin palestrën si një mënyrë për të qenë të shëndetshëm në mungesë të aktivitetit fizik të përditshëm. Këtë e bëjnë për të "shpëtuar " nga obeziteti dhe mbipesha, të dyja probleme të shpeshta. Gjithashtu, të gjithë e dinë, se nëse nuk ushtrohesh, muskujt dobësohen. Por, ajo që shumica e njerëzve nuk u kupton është që, dhe truri qendron në një gjendje më të mirë kur bën ushtrime fizike. Pra, ushtrimet fizike ndikojnë pozitivisht në të mësuarit dhe memorje.

Kjo është vënë re përmes një dekade studime në kafshë dhe njerëz.

Një nga provat e para për ndryshimet që sjellin ushtrimet në tru, është bërë që në vitet 1990. Dihej se truri kishte efekt në sjellje, por më parë askush nuk kishte pyetur për të kundërtën, e cila doli të ishte e vërtetë. Kur një grup shkencëtarësh të drejtuar nga Fred Gage, një neurolog nga Instituti Salk i Studime Biologjike, vendosi të krahasonte trurin e minjve të cilët kishin mundësi te pakufizuar për të lozur në një rrotë , ' minj vrapues' me ate të minjve' jovrapues'. Minjtë aktiv fizikisht kishin dyfishin e qelizave të reja nervore në hipokampus. Hipokampusi njihet si pjesa e trurit e përfshirë në aftësinë e të mësuarit dhe memorje. Eksperimentet treguan se minjtë e aftë fizikisht ishin më mirë në testet e memorjes. [37]

Sipas rezultateve, neurologia Judy Cameron[37] e Universitetit të Pittsburgh, shtroi pyetjen nëse ushtrimet fizike do të ndikonin edhe tek majmunët. Në laboratorin e saj u trajnuan një grup majmunësh të një moshe jo të re, në pista vrapimi speciale një orë cdo dite, pse ditë në javë për 5 muaj. Ndërkohë, një grup tjetër majmunësh nuk u përfshi në ushtrime. Ky eksperiment u përpoq të hidhte një vështrim te ndikimi i ushtrimeve mbi njerëzit, të cilët nuk janë si minjtë që vrapojnë vazhdimisht. Rezultatet treguan se majmunët në regjimin e caktuar, arriten të mësonin dyfish më shpejt se të tjerët.
Studimet mbi njerëzit, si ato të kryera nga Art Kramer, i cili vëzhgoi së si palestra mund të ndryshonte trurin gjatë plakjes, në Universitetin e Illinois, kanë treguar një lidhje të qartë mes të ushtruarit dhe përmirësimit të aftësive memorizuese gjatë jetës.

Sipas kërkimeve më të fundit, edhe njerëzit të cilët nuk kanë qenë shumë aktiv mund të përfitojnë nga aktivitetet , ndonëse në një moshe të thyer.

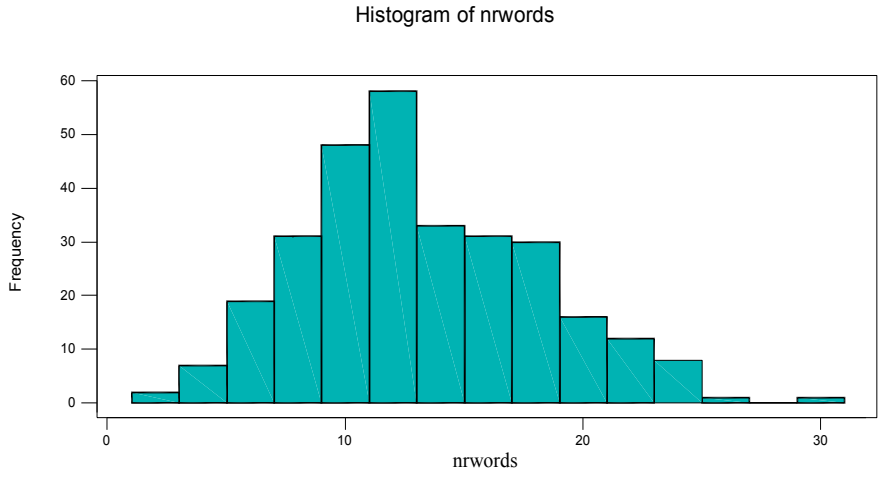
7.2 Cfarë e shkakton këtë rritje në memorje?

Ushtrimet fizike zmadhojnë volumin e gjakut në tru, duke rritur sasinë e oksigjenit të pompuar në cdo qelizë neuroniane.
Studimet kanë treguar që më shumë aktivitet fizik con në një nivel me të larte të molekules BDNF[38] (molekule kjo e njohur për vetinë neurotropike të trurit).
Një numër më i madh qelizash në rritje në hipokampus dhe një rritje në vëllim janë vënë re në subjekte që ushtrohen. [37, 40]
Një studim i bërë në Irland[38] zbuloi nëpërmjet analizave të gjakut se, pas ushtrimeve të rregullta, një nivel më i lartë BDNF, protein, gjendej në trup.
BDNF është një molekulë e familjes neurotropike. Kjo familje proteinash është e rëndësishme për mbijetesën , zhvillimin e funksionimin e neuroneve. Molekula e parë e ë familje NGF (Nerve Growth Factor) u zbulua nga fituesi i Cmimit Nobel, Levi Montalcini, dhe Hamburger më 1950. BDNF, pjestari i dytë i ë familje u zbulua në 1982. Më vonë janë zbuluar dhe të tjera proteina të ngjashme. [39]
BDNF, më e studiuara e llojit të saj, është e rëndësishme për zhvillimin e trurit gjatë embriogjenezës, fëmijërisë dhe rritjes. Kjo proteinë ndihmon mbijetesën e neuroneve duke ndikuar në rritjen, maturimin, dhe mirëmbajtjen e këtyre qelizave. Në tru, proteina e BDNF është aktive tek sinapses. Ajo rregullon plasticitetin sinaptik që është i rëndësishëm për procesin e të mësuarit dhe të memorizimit.
Duke patur parasysh nivelin e larte te BDNF në gjakun e njerëzve me nje jetesë të shendetshme, shkencëtaret besojnë se aktiviteti fizik permirëson memorjen dhe vëmendjen. [3, 4]
Një studim i publikuar online 31 January 2014, nga Akademia Kombëtare Shkencore[40] në zbuloi se ushtrimet rrisin volumin e hipokampusit te të rriturit nga 55 në 80 vjec, respektivisht me 2.12 në anën e majtë dhe 1.97 në të djathetën.Testet e memorjes mbi ta kanë treguar dhe një përmirësim të kësaj aftësie. [41,42]
Si konkluzion ushtrimet rrisin rrjedhjen e gjakut në tru dhe pompimin e oksigjenit. Në prezencën e nje sasie të mjaftueshme oksigjeni më shumë neurone rriten në hapësirën e hipokampusit . Bazuar në kërkimet e mësipërme, me këtë studim është synuar të provohet se rrezultatet e vëna re te të rriturit janë të njëjtë edhe te të rinjtë dhe adolishentët.

7.3 Metoda statistikore e ndjekur e interpretimi i rezultateve

Janë intervistuar rreth 300 persona të moshës 15-19 vjecare nga shkolla të ndryshme të Shqipërisë; kryesisht nga Tirana e rrethinat e saj, nga Durrësi, Shkodra, Gjirokastra e Vlora. Para eksperimentit subjekteve iu shpjegua studimi e procedura. Personat që kishin probleme dhe çrregullime në memorje, humbje të mëdha në peshë, gjendje shëndetësore shumë të keqe nuk u përfshinë në studim.
Pyetsori përfshinte informacione demografike për moshën, gjininë dhe qytetin e lindjes, si dhe pyetje për mënyrën e jetesës. Pyetjet për mënyrën e jetesës ishin pyetje për konsumimin e alkolit, traditën e të ngrënit të ushqimeve shumë të rënda ose frutave.
Për të përcaktuar memorjen e ndërmjetme (the intermediate term memory) është bërë një test për memorjen. Subjekteve i është dhënë një listë që përmban 40 fjalë të fushave të ndryshme, dhe ato janë lexuar për 5 minuta. Pastaj i është kërkuar të shkruanin fjalët që kanë mbajtur mend. Qëllimi i studimit ishte numri total i fjalëve të memorizuara.
Sasia e ushtrimeve fizike ishte e vështirë për tu matur, sepse ushtrime të ndryshme fizike ndryshojnë për nga koha që konsumohet për to, dhe për nga energjia. Kështu të luash tenis dhe të luash futboll nuk kërkon të njëjtin intesitet, por ato u maten njëlloj. Kjo do të thotë që *shpeshtësia në ushtrime* përmban sa 30 minuta ushtrime kryen gjatë një jave.
'Të ngënit shëndetshëm' ishte e vështirë gjithashtu të përcaktohej, kjo sepse personat konsumonin fruta psh, pas darke, së bashku me anëtarët e tjerë të familjes, duke i ndarë ato me njëri -tjetrin. Kështu, ishte e vështirë të rikujtohej dhe të përcaktohej sasia e konsumimit të frutave. Kjo është arsyeja pse ky variabël u konsiderua si variabël binar.
Duke shfrytëzuar këtë bazë të dhënash, u përcaktuan disa variabla. Variabli *nrfjalëve* , i cili përcaktohet nga numri i fjalëve të memorizuara për cdo person.

Figura 7–1 Histograma e numrit të fjalëve të memorizuara



U vu re se vlera më e vogël e tij ishte 1 dhe ajo më e larta 30. Më poshtë , Figura 7-1, paraqitet histogramën e këtij variabli me shkallë të zgjedhur 2 njësi. Derisa vëllimi i zgjedhjes ka qenë 300, pra i madh[43], për kontrollin e normalitetit të të dhënave të këtij variabli është përdorur testi i normalitetit Shapiro-Wilky, i cili tregon për jo normalitet të të dhënave, tabela 7-1 dhe 7-2.

Tabela 7–1 Analiza Deskriptive e nrword

Descriptives

		Statistic	Std. Error
nr total I fjaleve	Mean	14.144	.2966
	95% Confidence Interval for Mean — Lower Bound	13.561	
	95% Confidence Interval for Mean — Upper Bound	14.728	
	5% Trimmed Mean	13.986	
	Median	13.000	
	Variance	26.218	
	Std. Deviation	5.1204	
	Minimum	2.0	
	Maximum	33.0	
	Range	31.0	
	Interquartile Range	7.0	
	Skewness	.529	.141
	Kurtosis	.083	.281

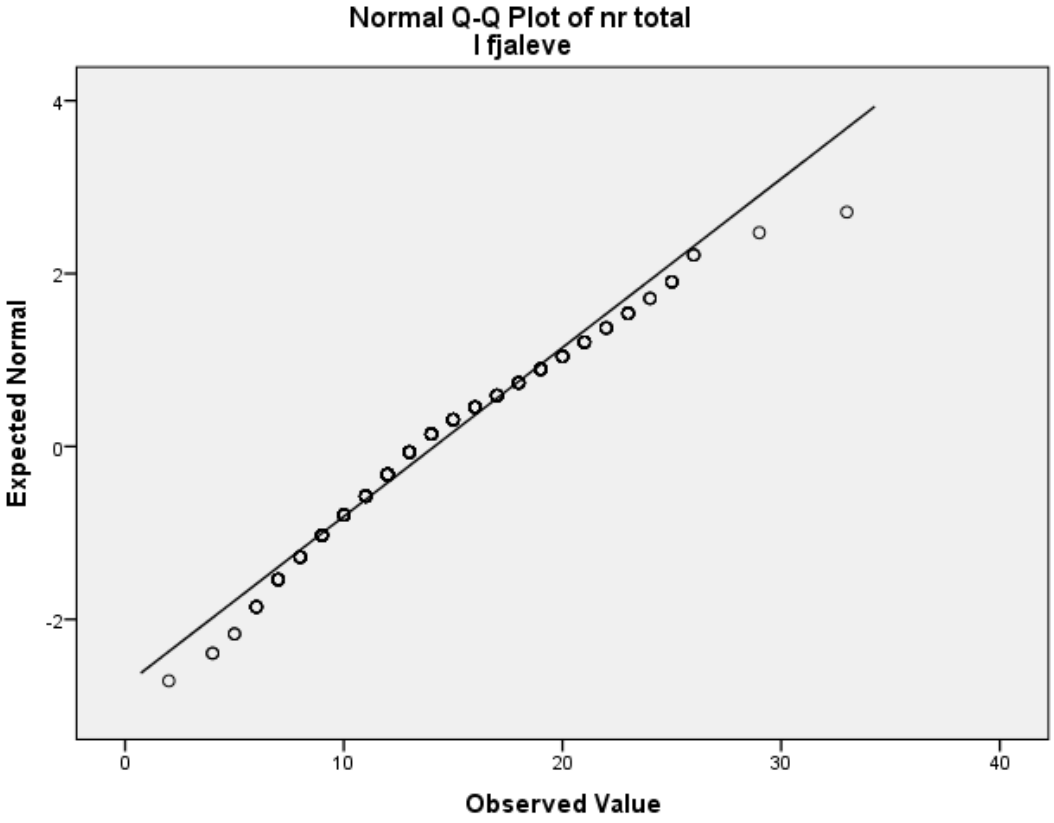
Tabela 7–2 Testi i Normalitetit të nrword

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
nr total I fjaleve	.109	298	.000	.976	298	.000

a. Lilliefors Significance Correction

Figura 7–2 Grafiku i ndryshimit të vlerave të *nrfjalëve* nga vlerat normale



Histograma tregon se kemi një pozitiv, ndaj punohet për normalizimin e të dhënave.

Transformimi me anë të funksionit *sqrt* të variablit në fjalë, krijon variablin e ri *sqrtnrfalëve*, i cili përmbush kushtet e normalitetit. Vlera e Skewness .018 dhe e Kurtosis -.058 janë tashme shumë pranë 0. Rezultati i Shapiro-Wilky, Table 7-3, tregon se variabli i transformuar ka shpërndarje normale. Ndaj, në analizen e mëvonshme mund të përdorim këtë variabël të ri.

Tabela 7–3 Analiza Deskriptive dhe testi i normalitetit të sqrt nrword

Descriptives

		Statistic	Std. Error
sqrtnword	Mean	3.6981	.03971
	95% Confidence Interval for Mean — Lower Bound	3.6199	
	95% Confidence Interval for Mean — Upper Bound	3.7762	
	5% Trimmed Mean	3.6990	

	Median	3.6056	
	Variance	.470	
	Std. Deviation	.68555	
	Minimum	1.41	
	Maximum	5.74	
	Range	4.33	
	Interquartile Range	.93	
	Skewness	.018	.141
	Kurtosis	-.058	.281

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
sqrtnword	.074	298	.000	.993	298	.154

a. Lilliefors Significance Correction

Variablat si *gjinia* dhe ato që kanë të bëjnë me mënyrën e jetesës, si *konsumi i duhanit*, *konsumi i alkolit* dhe *numri i ushtrimeve fizike gjatë javës* , pas analizës së korelacionit mund të konsiderohen të pavarura.

Të studiosh rolin e *duhanit* , të *të ngënit*, dhe *shpeshtësisë së ushtrimeve* në *nrfjalëve*, është përdorur metoda e regresit stepwise[44]. Rezultate e Tabelës 7-4 tregojnë se faktorë kryesorë në këtë modelim është *numri i ushtrimeve fizike gjatë javës* dhe *konsumi i duhanit*. Modeli në fjalë përcakton 40% të ndryshimeve të variablit në shqyrtim. Testi qe R=0, pra nuk ka lidhje lineare mes modelit te propozuar dhe faktoreve te tjerë te testuar, bie poshte eshte signifikative me $p<0.001$ Madje R=0.612 tregon se lidhja ne fjale eshte e fortë.

Figura 7–3 Grafiku i ndryshimit të vlerave të *sqrtnrfjalë ve* nga vlerat normale

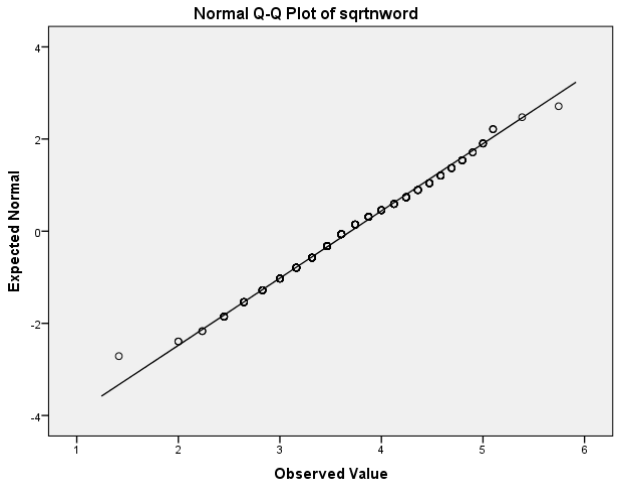


Tabela 7–4 Rezultatet e stepwise regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	saherebejnes portnejave	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	duhan po/jo	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: sqrtnword

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.624[a]	.389	.387	.53684
2	.639[b]	.409	.405	.52887

a. Predictors: (Constant), saherebejnesportnejave

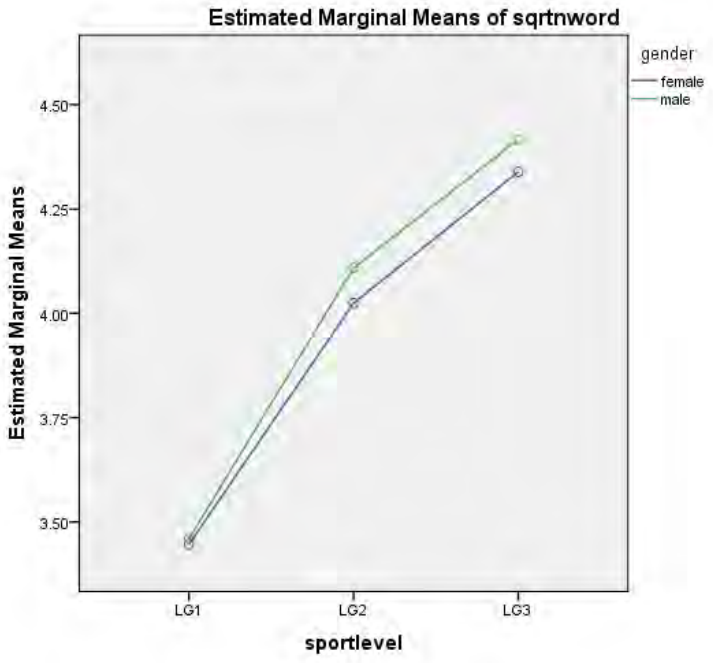
b. Predictors: (Constant), saherebejnesportnejave, duhan po/jo

Mesatarja e *nrfalëve* kur subjektet nuk konsumojnë duhan është 13.270, kurse vlera mesatare e *nrfjalëve* kur të personat konsumojnë duhan është më e ulët, 11.714. Vlerat mesatare janë përkatësisht 12.701 dhe 12.506 për personat që hanë shëndetshëm dhe jo. Por nga regresi stepwise këto variabla, *alkol* dhe *të ngrënit shëndetshëm* nuk janë marrë parasysh, pra nga baza e të dhënave nuk kemi ndikim të tyre në variablin *sqrtnrfjalëve*.
Bazuar në të dhënat në fjalë fitet për ndikim të *alkolit* dhe *numri i ushtrimeve fizike gjatë javës*.
However the two last results of this investigations are in contrast with the research "Relationship of serum brain-derived neurotrophic factor (BDNF) and
health-related lifestyle in healthy human subjects" by Ka Lok Chan, which reports higher BDNF level when the subjects eat healthy and when they smoke.
Por duhet thënë se këto dy rezultate bien në kundërshtim me rezultatet e punimit të Ka Lok Chan, "Lidhja e BDNF me mënyrën e jetesës së subjektit", në të cilen konkludohet se niveli i BDNF është më i latë nëse subjektet ushqehen më shëndetshëm dhe kur konsumojnë duhan.
Studimi i rolit të duhanit në *nrfjalëve,* tregoi se 212 subjekte që nuk konsumojnë duhan kanë një mesatare të *nrfjalëve* prej 13.182. Ndërsa 85 njerëz të cilët konsumojnë duhan kanë një mesatare më të ulët të *nrfjalëve* , prej 11.667 fjalësh. Koeficienti -.202 tregon për një rol negativ të konsumit të *duhanit* .
Ndërsa keficienti tjetër .230 pranë *numri i ushtrimeve fizike* tregon që mesatarisht numri i flalëve të memorizuara rritet nëse numri i orëve të ushtrimeve fizike rritet.
Nëqoftëse numri i ushtrime fizike rritet, nga dy herë në javë (sic është bërë zakonisht në shkollat tona) bëhet tre, atëherë numri i fjalëve të memorizuara ndryshon, nga 11.7 në 13.6.
Studimi u përqëndrua në 3 variabla: *gjini*, *numri i ushtrimeve fizike* dhe *sqrtnrfjalë*.

Në këtë databazë për përcaktimin e gjinisë është përdorur kjo mënyrë: kodi 1=femra dhe kodi 2= meshkuj. Kështu nga 300 të intervistuar të të dy gjinive; 178 ishin femra (kodi 1) dhe 119 meshkuj (kodi 2).
Duke patur parasysh se shumica e adoleshentëve në Shqipëri bëjnë dy herë në javë ushtrime fizike, subjektet janë grupuar në tre grupe kryesore:
Grupi i nivelit 1 (LG1) 0-2 herë ushtrime fizike gjatë javës
Grupi i nivelit 2 (LG2) 3-4 herë ushtrime fizike në javë
Grupi i nivelit 3 (LG3) më shumë se 4 herë ushtrime fizike në jave.

ANOVA Dypërmasore[9] është përdorur për të shqyrtuar ndikimin apo jo të variablave *gjini* me dy vlera , femra e meshkuj, dhe *grupi i nivelit* të ushtrimeve fizike, si variabel me tri vlera, në sqrtnrfjalëve. Grafiku i rezultateve, Figura 7-4, tregon se mesatarja e *sqrtnrfjalëve* është më e lartë në të tre grupet e niveleve për femrat se sa për meshkujt.

Figura 7–4 Grafiku i mesatareve të *sqrtnrfjalëve* sipas *gjinise* dhe *grupit te nivelit* te shtrimeve fizike



Por rezultatet e Tabelës 7-5 (F .333 dhe Sig .565) tregojnë qartë se gjinia nuk sjell asnje diferencë në *sqrtnrfjalëve* për cdo grup të nivelit të sportit.

Tabela 7–5 Rezultatet e ANOVA dy përmasore

Tests of Between-Subjects Effects

Dependent Variable: sqrtnword

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared

Corrected Model	37.779[a]	5	7.556	21.672	.000	.271
Intercept	2148.261	1	2148.261	6161.623	.000	.955
gender	.116	1	.116	.333	.565	.001
sportlevel	29.465	2	14.733	42.256	.000	.224
gender * sportlevel	.072	2	.036	.103	.902	.001
Error	101.806	292	.349			
Total	4215.000	298				
Corrected Total	139.585	297				

a. R Squared = .271 (Adjusted R Squared = .258)

Duke u përqëndruar në afishimet pranë variablit nivel i sportit (F=42.256 dhe Sig. 7.62 E-17), kuptohet lehtë se *sqrtnrfjalëve* rritet ndersa numri i ditëve të ushtrimeve fizike rritet.
F =.103 and Sig.902 , tregojnë se nuk ka nderhyrje mes dy faktorëve në fjalë.
Përqindjet e tre grupeve të ushtrimeve fizike sipas gjinise, tregojnë se femrat përbëjnë 73.91 % të LG1, ndërsa meshkujt 72.34% të LG3. Pra, shifrat se meshkujt kanë, në të memorizuarën e fjalëve, një mesatare më të lartë se femrat, lidhet me atë që djemtë bejnë më shumë ushtrime fizike se femrat.
Pas kësaj, nuk është e nevojshme të kontrollohet lidhja mes *gjinisë* e numrit të fjalëve të memorizuara, por vetëm ndikimi i *numrit të ditëve të ushtrimeve fizike* në *sqrnrfjalë.*
Me metodën e krahasimit cift e cift, Tabela 7-6, është testuar se, kur ndryshimet mes dy niveleve të njëpasnjëshme të variablit *sportlevel* ,janë statistikisht të dukshëme. Rritja, nëse përqëndrohesh në krahasimin e dy grupeve të njëpasnjëshme, është e ndjeshme kur bëhet kalimi nga Grupi i Nivelit 1 të sportit në Grupi i Nivelit 2(Sig 3.8 E-12). Ndërsa mes dy grupeve të fundit s' është statistikisht e ndjeshme me nivel rendesie α=0.01.

Tabela 7–6 Rezultatet e krahasimit të sqrtnrfjalëve sipas grupeve të nivelit të sportit

Pairwise Comparisons

Dependent Variable: sqrtnword

(I) sportlevel	(J) sportlevel	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
0-2 here sport	3- 4 here sport	-.616*	.082	.000	-.778	-.454
	me shume se 4 here sport	-.950*	.111	.000	-1.167	-.732
3- 4 here sport	0-2 here sport	.616*	.082	.000	.454	.778
	me shume se 4 here sport	-.334*	.124	.0077	-.579	-.089
me shume se 4 here sport	0-2 here sport	.950*	.111	.000	.732	1.167
	3- 4 here sport	.334*	.124	.0077	.089	.579

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Nuk ka arsye të mos besojme se, rritja e numrit të ushtrimeve fizike rrit numrin e fjalëve të memorizuara. Për më tepër, nese një adoleshent bën një orë më shumë ushtrime fizikese sa bëhej zakonisht në Shqipëri, e kalohet nga LG1 në LG2, atëhehe pritet të ketë statistikisht aftësi më të larta memorizuese.

7.4 Përfundime

Bazuar në rezultatet e mësipërme mund të themi:

- Ka një ndikim të ushtrimeve fizike në memorjen e adoleshentëve. Rritja e sasisë së ushtrimeve fizike, rrit shkallën e memorjes tek adoleshentët.
- Nuk ka dallim statistikor në aftësitë memorizuese mes djemve e vajzave.
- Ka një ndikim të konsumimit të duhanit në memorje.
- Por, bazuar në këtë studim, s'mund të thuhet se të ngrënit shëndetshëm dhe alkoli kanë ndikim në aftësitë memorizuese të adoleshentëve.

Referencat

[1] Scheffé, H. (1959). *The Analysis of Variance*. Wiley, NeW York. (reprinted 1999,ISBN 0-471-34505-9), pg. 154-197

[2]ABDI & WILLIAMS, Analysis of contrast/www.utd.edu/~herve/abdi-contrasts2010-pretty.pdf .(2010)

[3] Afifi A.A., Azen S.P. "Statistical analysis. Computer oriented approach", (1979): Academic press New York, San Francisco, London,pg.198-274.

[4] An introduction to Applied Multivariate Analysis, Tenko Raikov and George Marcoulides,99-150.

[5] Gunst R.F., Mason R.L. "Regression analysis and its application", Marcel Dekker Inc., NewYork (1980), pg.121-189.

[6] Afifi A.A., Azen S.P. "Statistical analysis. Computer oriented approach", (1979): Academic press New York, San Francisco, London, pg.124-197.

[7] Miller, I., Freund, J. E., (1978): Probability and Statistics for Engineers Prentice-Hall, Inc.,pg.261-310.

[8] IBM SPSS Regression 22, Chapter 2, Multiple Linear Regression, 2013.

[9] Miller, I., Freund, J. E., (1978): Probability and Statistics for Engineers Prentice-Hall, Inc.,pg.226-264.

[10] U.S. Department of Health and Human Services. (1998): Toxicological profile for sulphur dioxide. Public Health Service, Agency for Toxic Substances and Disease Registry.

[11] Scheffé, H. (1959). *The Analysis of Variance*. Wiley, NeW York. (reprinted 1999,ISBN 0-471-34505-9)

[12]ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/en/client/Manuals/IBM_SPSS_Statistics_Command_Syntax_Reference.pdf

[13] The Truth about Principal Components and Factor Analysis www.stat.cmu.edu/~cshalizi/350/lectures/13/lecture-13.pdf, (2009). (Bartholomew, David J. , 1987).

[14] Barbara G. Tabachnick and Linda S. Fidell, Using Multivariate Statistics, Third Edition. Logistic Regression chapter.(2012) .

[15] Afifi A.A., Azen S.P. “Statistical analysis. Computer oriented approach”, (1979): Academic press New York, San Francisco, London,pg.301-353.

[16] Bartholomew, David J. (1987). Latent Variable Models and Factor Analysis. New York: Oxford University Press. Stone, James V. (2004).

[17] Independent Component Analysis: A Tutorial Introduction. Cambridge, Massachusetts: MIT Press. Thomson,(1982),sec.2.

[18] Using Multivariate Statistics, 6/E Barbara G. Tabachnick,Linda S. Fidell, pg.215-278'

[19] Casella Berger Statistical Inference (2002) ,pg. 600-670

[20] https://mathdept.iut.ac.ir/sites/mathdept.iut.ac.ir/files/AGRESTI.PDF pg120-208

[21]Categorical Data Analysis, Alan Agresti, ISBN:0-471-36093-7, pg.590-630

[22] Categorical Data Analysis, Alan Agresti, ISBN:0-471-36093-7, seks3.

[23]Altman D.G. (1991) Practical Statistics for Medical Research. Chapman & Hall, www.blackwellpublishing.com/specialarticles/jcn_10_774.pdf

[24] London. Polit D. (1996) Data Analysis and Statistics for Nursing Research. Appleton & Lange, Stamford, pg.579-671.

[25] Categorical Data Analysis, Alan Agresti, ISBN:0-471-36093-7, pg.169-264.

[26]http://practicalstats.labanca.net/index.php?title=Rules_of_thumb_for_interpreting_the_size_of_a_correlation_coefficient

[27] IBM SPSS Regression 22, Chapter 2, Logistic Regression, 2013.

[28] Casella Berger Statistical Inference, second Edition , George Casella, Roger L. Berger,2002, pg. 600-670.

[29] Applied Logistic Regression , David W. Hosmer, Jr., Stanley Lemeshow, Rodney X. Sturdivant, pg.13-52.

[30] http://www.hcdoes.org/airquality/monitoring/so2.htm

[31] http://www.missouribotanicalgarden.org/gardens-gardening/your-garden/help-for-the-home-gardener/advice-tips-resources/pests-and-problems/environmental/sulfur-dioxide.aspx

[32] An introduction to Applied Multivariate Analysis, Tenko Raikov and George Marcoulides Chapter 9,391-427

[33] http://www.dokeefe.net/pub/okeefe07cmm-posthoc.pdf

[34]An introduction to Applied Multivariate Analysis, Tenko Raikov and George Marcoulides,210-276.

[35] Applied Longitudinal Analysis, G. Fitzmaurice, N. Laird and Ware, 490-513

[36] Practical Multivariate Analysis, Fifth Edition
By Abdelmonem Afifi, Susanne May, Virginia, pg. 231-264

[37] http://www.brainfacts.org/across-the-lifespan/diet-and-exercise/articles/2013/physical-exercise-beefs-up-the-brain/

[38] http://www.ncbi.nlm.nih.gov/pubmed/21722657
Griffin ÉW, et al. Aerobic exercise improves hippocampus function and increases BDNF in the serum of young adult males. Physiol Behav. 2011 104(5):934-41

[39] Devin K. Binder and Helen E. Scharfman. Brain-derived Neurotrophic Factor *Growth Factors*. 2004 September; 22(3): 123–131.

[40] Ka Lok Chan et al., Relationship of serum brain-derived neurotrophic factor (BDNF) and health-related lifestyle in healthy human subjects (The results of this study were approved by the Research Ethnic Committee of Hong Kong Hospital Authority and the Hong Kong Polytechnic University Human Subjects Ethnics Sub-Comittee)

[41] http://www.pnas.org/content/108/7/3017.short
Kirk I. Erickson et al., Exercise training increases size of hippocampus and improves memory, PNAS February 15, 2011 vol. 108 no. 7 3017–3022

[42] Kirk I. Erickson et al Brain-Derived Neurotrophic Factor Is Associated with Age-Related Decline in Hippocampal Volume The Journal of Neuroscience, April 14, 2010 • 30(15):5368 –5375

[43] Applied Longitudinal Analysis, G. Fitzmaurice, N. Laird and Ware, *490-513*

[44] Practical Multivariate Analysis, A.A. Afifi , S.May, V. Clarc (2011), 231-264

[45] An introduction to applied multivariate analysis, Tenko Raikov and George Marcoulides (2008),*120-314*