Tampereen yliopisto

Juulia Kärki

Kovariaattien merkitys ANCOVA-mallissa

Informaatioteknologian ja viestinnän tiedekunta
Tilastollisen data-analyysin kandidaattitutkielma
Huhtikuu 2021

Tiivistelmä

Juulia Kärki: Kovariaattien merkitys ANCOVA-mallissa
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Huhtikuu 2021

Kovarianssianalyysi (*Analysis of covariance*) on yleinen lineaarinen malli, joka on yhdistelmä regressio- ja varianssianalyysista. Kovarianssianalyysia käytetään tutkimaan jatkuvien ja luokiteltujen muuttujien vaikutusta selitettävään muuttujaan. Siinä voidaan käyttää yhtä tai useampaa selittävää muuttujaa, joita kutsutaan kovariaateiksi.

Tämä tutkielma voidaan jakaa kahteen osaan. Ensimmäisessä osassa esitellään teoreettisesti kovariaattianalyysin malli, sen tärkeimpien parametrien estimointi, hypoteesin testaus ja kovarianssianalyysin taulu. Tärkeimmät analyysin tunnusluvut ovat esitetty taulussa. Toisessa osassa hyödynnetään teoriaosuutta käytännön aineistoon. Testauksessa käytetään Suomen Ekonomit järjestön keräämää palkkatasoaineistoa ja hyödynnetään sitä kovarianssianalyysin testaukseen. Ensimmäiseksi esitellään testauksessa käytettyjen muuttujien jakautumista havaintoaineistossa. Testauksessa tarkastellaan kovariaattien merkitystä kovarianssianalyysissa ja tutkitaan erilaisten muuttujien vaikutuksia testituloksiin.

Havaintoaineistosta käytetään ryhmiinjakokriteereinä sukupuolta ja selittävänä muuttujana eli kovariaattina ikää ja yrityksen henkilöstömäärää tutkimaan selitettävän muuttujan palkan jakautumista. Näistä kovariaateista tehdään kolme erilaista kovarianssianalyysin mallia, esitellään niiden parametriestimaatit ja testataan estimoidun kovarianssianalyysin mallia. Näistä kaikista malleista tutkitaan kovariaattien merkitys mallissa ja tutkitaan residuaalien käyttäytymistä graafisesti. Kovarianssianalyysin testaukset ja graafiset esitykset ovat tehty R-ohjelmistokielen avulla.

Ensimmäiseen malliin lisätään kovariaatiksi ikämuuttuja. Palkkoissa on eroavaisuuksia, kun iän vaikutus on poistettu. Toisessa mallissa korvataan kovariaatti ikämuuttuja henkilöstömäärämuuttujalla. Henkilöstömäärän vaikutuksen poisto ei vaikuta palkkaeroihin tilastollisesti merkitsevästi. Kolmanteen malliin on sisällytetty kumpikin kovariaatti ikä ja henkilöstömäärä. Kun kovariaattien iän ja henkilöstömäärän vaikutus on poistettu, palkoissa on eroavaisuuksia sukupuolittain. Kovariaattien lisääminen malliin ei siis juurikaan vähentänyt palkkaeroja. Miesten ja naisten palkoissa on eroa, mutta palkkojen eroa ei selitä henkilöstömäärä ja ikä. Palkkaeroon vaikuttavat muut tekijät enemmän.

Avainsanat: kovarianssianalyysi, varianssianalyysi, parametrien estimointi,
harhaton estimaatti, jäännösneliösumma

Sisältö

1 Johdanto

Kovarianssianalyysi perustuu yleiseen lineaariseen malliin. Sitä käytetään usein lääketieteessä tilastolliseen testaamiseen. Tämä menetelmä on niin sanotusti regressio- ja varianssianalyysin sekoite. Analyysin tarkoituksena on tutkia jatkuvien ja luokiteltujen muuttujien vaikutusta riippuvaan muuttujaan. Menetelmän avulla voidaan havaita muuttujien välillä merkitsevyyttä, vaikka sitä ei olisi havaittu varianssianalyysin tapauksessa. Tässä tutkielmassa tarkastellaan kovarianssianalyysin mallia, ja sovelletaan sitä minimoimaan tutkielmassa käytetyn aineiston kovariaattien sekoittavaa vaikutusta pois.

Vaikka kovarianssianalyysia on käytetty jo vuosia, se on edelleen ehkä yksi vähiten ymmärretty ja eniten väärinkäytetty tilastollinen menetelmä. Jopa pätevät tutkijat käyttävät mallia tilanteissa, joissa tällainen analyysi johtaa virheellisiin johtopäätöksiin, eikä mallia käytetä silloin, kun se olisi tarkoituksenmukaista. (Huitema, 2011, XV.)

Tutkielmassa esitellään kovarianssianalyysin teoreettista taustaa luvussa 2. Teoriaosuudessa esitellään yleinen kovarianssianalyysin malli, parametrien estimointi, hypoteesin testaus ja esitellään ANCOVA-taulu.

Luvussa 3 testataan erilaisilla kovarianssimalleilla muuttujien riippuvuutta tutkielmassa olevan esimerkkiaineiston avulla. Tutkitaan keskimääräistä palkkaa sukupuolittain, kun taustakovariaattien vaikutus on vakioitu. Kovariaatteina käytetään ikää ja yrityksen henkilöstömäärää. Näistä kovariaateista muodostetaan kolme erilaista mallia. Ensimmäisessä mallissa käytetään ikämuuttujaa kovariaattina, toisessa mallissa henkilöstömäärämuuttujaa kovariaattina ja kolmannessa mallissa molemmat muuttujat ovat sisällytetty malliin. Näistä malleista tehdään kovarianssianalyysi ja tarkastellaan näiden muuttujien merkitystä kovariaatteina.

Tutkielman lukijalta oletetaan tilastollisten menetelmien, ja matriisilaskennan tietämystä. Lisäksi lukijan oletetaan ymmärtävän regressio- ja varianssianalyysin perusteet.

2 Tilastolliset menetelmät

Tämän luku pohjautuu pääasiassa Debasis S. ja Screenivasa R.J. kirjan *Linear models an integrated approach* lukuun 6.6.

Opinnäytetyössä sovelletaan kovarianssianalyysia (*ANCOVA*), joka yhdistää varianssianalyysin ja lineaarisen regressioanalyysin mallit. Menetelmällä voidaan tutkia, ovatko ryhmien väliset erot tilastollisesti merkitseviä, kun jatkuva muuttujan vaikutusta pyritään kontrolloimaan. Kovarianssianalyysi on yleinen lineaarinen malli, joka sisältää yhden selitettävän jatkuvan muuttujan, vähintään yhden selittävän kategorisen muuttujan (*faktorin*) sekä vähintään yhden jatkuvan muuttujan, jota kutsutaan kovariaatiksi. Analyysi on saanut nimensä tästä kovariaatista. (Korner-Nievergelt, 2015, s.56.) Tässä luvussa esitellään malli, sen parametrien estimointi ja hypoteesin testaus.

2.1 Kovarianssianalyysin malli

Tarkastellaan ensimmäiseksi yleistä kovarianssianalyysin mallia. Malli voidaan esittää matriisimuodossa:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{2.1}$$

missä

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{12} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\boldsymbol{Z} = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1q} \\ z_{21} & z_{22} & \dots & z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{nq} \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_q \end{pmatrix} \quad \text{ja} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Matriisimuodossa on esitetty $\boldsymbol{y}$ havaintojen y_i satunnainen vektori (n x 1) selitettävästä muuttujasta, matriisi $\boldsymbol{X}$ on suunnittelumatriisi (n x p), jossa x_{ij} ovat indikaattori muuttujia (0 tai 1), jotka sisältävät ryhmiin jakokriteerit. Vektorin $\boldsymbol{\beta}$ parametrit β_j ovat ryhmäkeskiarvoja, matriisi $\boldsymbol{Z}$ sisältää kovariaatit sekä $\boldsymbol{\eta}$ on kovariaattien kertoimien vektori (q x 1). Vektori $\boldsymbol{\epsilon}$ on virhetermien ϵ_i vektori (n x 1). Indeksit $i = 1, \dots, n$ ja $j = 1, \dots, p$. Virhetermi noudattaa normaalijakaumaa $E(\boldsymbol{\epsilon}) = \boldsymbol{0}, Var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ (Sengupta & Jammalamadaka, 2003, s.224).

Edellä mainittu kovarianssianalyysin malli (2.1) on yleistys mallista

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

missä x_1 sisältää ryhmiinjakokriteerit, x_2 on kovariaatti, β_0, β_1 ja β_2 ovat estimoitavia parametreja ja ϵ on satunnaisvirhetermi. Muuttujan x_2 sisällyttäminen malliin on tarkoitus vähentää selittämättömien tekijöiden vaikutusta. (Sengupta & Jammalamadaka, 2003, s.3)

2.2 Mallin käyttö

Kovarianssianalyysin avulla voidaan tutkia samanaikaisesti luokitteluasteikollisten muuttujien eli indikaattorimuuttujien ja jatkuvien muuttujien vaikutusta selitettävään muuttujaan. Analyysia voidaan käyttää sekä havainnoivissa tutkimuksissa, että kokeellisessa tutkimuksessa. Perusajatuksena on tutkia jatkuvan muuttujan ja indikaattorimuuttujan vaikutusta ryhmien vertailussa varianssianalyysilla. Muuttujien lisäyksen avulla olisi tarkoitus vähentää mallin virhetermien varianssia. Kovariaatti on olemassa yhdessä tai useammassa apumuuttujassa.

Pyritään pienentämään sekoittavien tekijöiden eli kovariaattien vaikutusta. Joissakin tapauksissa ongelma on arvioitaessa faktorien vaikutusta, jotka ovat merkitsevästi erilaisia, kun kovariaattien vaikutusta pienennetään. Mallin avulla voidaan pienentää koevirhettä.

2.3 Parametrien estimointi

Tämän luvun parametrien estimointien merkinnät pohjautuvat enimmäkseen Debasis S. ja Screenivasa R.J. kirjan *Linear models an integrated approach* lukuun 6.6.3 *Estimation of parameters*.

Aluksi tarkoituksena on hyödyntää suunnittelumatriisien $\boldsymbol{X}$ ja $\boldsymbol{Z}$ rakennetta parametrien $\boldsymbol{\eta}$ ja $\boldsymbol{\beta}$ estimoinnissa. Parametri $\boldsymbol{\eta}$ on otettava huomioon analysoidessa niitä parametreja, joista ollaan kiinnostuneita, vaikka ensisijainen huomio ei olekaan parametreissa $\boldsymbol{\eta}$. Tällaista parametria kutsutaan haittaparametriksi (*nuisance parameters*).

Mallista 2.1 voidaan muodostaa normaaliyhtälö parametreille $\boldsymbol{\beta}$ ja $\boldsymbol{\eta}$

$$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\eta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{pmatrix}. \tag{2.2}$$

Olkoon $(\boldsymbol{X'X})^{-1}$ matriisin $\boldsymbol{X'X}$ käänteismatriisi, ja korvataan $\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}$. Täten yhtälöstä 2.2 saadaan $\boldsymbol{\beta}$ estimaattiksi

$$\widehat{\beta} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'y} - \boldsymbol{X'Z}\widehat{\boldsymbol{\eta}}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\eta}}). \tag{2.3}$$

Estimoitu parametri $\boldsymbol{\eta}$ saadaan

$$\widehat{\boldsymbol{\eta}} = (\boldsymbol{Z'}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'})\boldsymbol{Z})^{-1}\boldsymbol{Z'}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'})\boldsymbol{y}. \tag{2.4}$$

Sijoitetaan $\boldsymbol{P}_x = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ yhtälöön 2.4, josta saadaan parametrille $\boldsymbol{\eta}$ estimaatti

$$\widehat{\boldsymbol{\eta}} = (\boldsymbol{Z'}(\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{Z})^{-1}\boldsymbol{Z'}(\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{y}.$$

Estimoitu parametri $\boldsymbol{\eta}$ voidaan merkitä lyhyesti $\widehat{\boldsymbol{\eta}} = \boldsymbol{R}^{-1}\boldsymbol{r}$, missä $\boldsymbol{R} = \boldsymbol{Z'}(\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{Z}$ ja $\boldsymbol{r} = \boldsymbol{Z'}(\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{y}$.

Paras lineaarinen harhaton estimaattori (*BLUE*) jollekin $\boldsymbol{A\beta}$ estimoituvalle funktiolle on muotoa $\boldsymbol{A\widehat{\beta}}$, missä $\widehat{\boldsymbol{\beta}}$ on kuten kohdassa (2.3) ja $\widehat{\boldsymbol{\eta}}$ on kohdassa (2.4).

Huomataan, että estimaatin $\widehat{\boldsymbol{\beta}}$ lauseketta voidaan tulkita seuraavalla tavalla

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{Z\eta} + \boldsymbol{\epsilon} \\ &= \boldsymbol{X\beta} + \boldsymbol{P}_X\boldsymbol{Z\eta} + (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{Z\eta} + \boldsymbol{\epsilon} \\ &= \boldsymbol{X\beta_0} + (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{Z\eta} + \boldsymbol{\epsilon}, \end{aligned} \tag{2.5}$$

missä $\boldsymbol{\beta}_0 = \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\eta}$. Olkoon $\boldsymbol{Z}_j$ matriisin $\boldsymbol{Z}$ sarake, joka sisältää $j = 1, \ldots, q$ kappaletta kovariattien muuttujia. Täten voidaan kirjoittaa

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 - \sum_{j=1}^{q} \eta_j \widehat{\boldsymbol{\alpha}}_j,$$

missä η_j on parametrin $\boldsymbol{\eta}$ elementit j:hin asti ja $\widehat{\boldsymbol{\alpha}_j}$ on pienimmän neliösumman estimaattori $\boldsymbol{\alpha}_j$ mallista $(\boldsymbol{z}_j, \boldsymbol{X\alpha}_j, \sigma^2\boldsymbol{I})$. Korvaamalla parametrit $\boldsymbol{\beta}_0$ ja $\boldsymbol{\eta}$ niiden pienimmän neliösumman estimaattorilla $\widehat{\boldsymbol{\beta}_0} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Zy}$ sekä $\widehat{\boldsymbol{\eta}}$ kohdan 2.4 mukaan, saadaan

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}_0} - \sum_{j=1}^{q} \widehat{\eta}_j \widehat{\boldsymbol{\alpha}}_j.$$

Tämä lauseke vastaa estimaattia 2.3, kun malli 2.5 uudelleen parametrisoidaan mallin 2.1 avulla (Sengupta & Jammalamadaka, 2003, s.227).

2.4 Hypoteesin testaus

Oletetaan, että vektorin $\boldsymbol{y}$ ehdollinen jakauma ehdolla $\boldsymbol{Z}$ on normaalijakauma (Sengupta & Jammalamadaka, 2003, s.228). Tarkastellaan ensiksi hypoteesia $\boldsymbol{\eta} = \boldsymbol{0}$, jossa kovariaatit voidaan jättää huomioimatta. Hypoteesin kokonaisneliösumma SST on $R_H^2 = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}$ vapausastein $n - \rho(\boldsymbol{X})$, missä $\rho(\boldsymbol{X})$ on matriisin $\boldsymbol{X}$ aste (Sengupta & Jammalamadaka, 2003, s.228). Tällöin mallin 2.1 jäännösneliösumman on

$$SSE = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P}_{X:Z})\boldsymbol{y} = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P}_X - \boldsymbol{P}_{(I-P_X)Z})\boldsymbol{y}.$$

Hyödyntämällä malleja 2.3 ja 2.4 voidaan merkitä jäännösneliösummaa

$$SSE = \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X'y} - \widehat{\boldsymbol{\eta}}'\boldsymbol{Z'y}$$

vapausastein $n - \rho(\boldsymbol{X} : \boldsymbol{Z}) = n - \rho(\boldsymbol{X}) - \rho(\boldsymbol{Z})$, missä $\rho(\boldsymbol{Z})$ on matriisin $\boldsymbol{Z}$ aste. Tästä seuraa, että uskottavuusosamäärätesti (*GLRT*) nollahypoteesille $\boldsymbol{\eta} = \boldsymbol{0}$ voidaan hylätä, kun suhde $[(n-\rho(\boldsymbol{X} : \boldsymbol{Z}))(\boldsymbol{r'R}^{-1}\boldsymbol{r})/(\rho(\boldsymbol{X} : \boldsymbol{Z})-\rho(\boldsymbol{X}))SSE]$ on suuri. Nollahypoteesin vallittaessa testisuure noudattaa F-jakaumaa $F_{\rho(X:Z)-\rho(X),n-\rho(X:Z)}$. (Sengupta & Jammalamadaka, 2003, s.228) Lyhyemmin jäännösneliösummaa voidaan merkitä

$$SSE = R_H^2 - \boldsymbol{r'R}^{-1}\boldsymbol{r}. \tag{2.6}$$

Seuraavaksi tarkastellaan yleisen lineaarisen hypoteesin testausta ja sen päähaasteita, jossa hypoteesi on muotoa $\boldsymbol{A}\beta = \boldsymbol{0}$. Testauksessa hypoteesillä ei ole käsittelyvaikusta, lohkovaikutusta, interaktiota eikä sisäkkäistä vaikutusta. Nämä kaikki ovat tämän ongelman erikoistapauksia. Tästä seuraa, että malli 2.7 on yhtäpitävä

$$\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}_{A'})\boldsymbol{\theta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \qquad E(\boldsymbol{\epsilon}) = \boldsymbol{0} \qquad \text{ja} \qquad D(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}. \tag{2.7}$$

Hyödynnetään jäännösneliösummaa 2.6 mallin 2.7 jäännösneliösummalle. Tällöin jäännösneliösumma on

$$SSH = R^2_{H\beta} - \boldsymbol{r}'_H\boldsymbol{R}^{-1}_H\boldsymbol{r}_H, \tag{2.8}$$

missä jäännösneliösumman oikeanpuoleiset termit ovat saatu

$$\begin{pmatrix} R^2_{H\beta} & \boldsymbol{r}'_H \\ \boldsymbol{r}_H & \boldsymbol{R}_H \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}' \\ \boldsymbol{Z}' \end{pmatrix} (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}(\boldsymbol{I}-\boldsymbol{P}_{A'})}) \begin{pmatrix} \boldsymbol{y} & \boldsymbol{Z} \end{pmatrix}.$$

Vapausasteet termille jäännösneliösummalle SSE on $n - \rho(\boldsymbol{X} : \boldsymbol{Z})$ ja vastaavasti vapausasteet jäännösneliösummalle SSH on $n - \rho(\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}_{A'}) : \boldsymbol{Z})$. Täten hypoteesin $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ GLRT voidaan hylätä tilastollisen testisuureen arvoilla

$$\frac{SSH - SSE}{SSE} \cdot \frac{n - \rho(\boldsymbol{X} : \boldsymbol{Z})}{(\boldsymbol{X} : \boldsymbol{Z}) - \rho(\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}_{A'}) : \boldsymbol{Z})}, \tag{2.9}$$

missä testisuureen arvo noudattaa $F_{\rho(X:Z)-\rho(X(I-P_{A'}):Z),n-\rho(X:Z)}$. Jos kovariaatit ovat riippumattomia erilaisista vaikutuksista, $C(\boldsymbol{Z})$ ja $C(\boldsymbol{X})$ ovat itse asiassa epäjatkuvia. (Sengupta & Jammalamadaka, 2003, s.229)

2.5 ANCOVA -taulu

Neliösummat on esitetty muodossa $R^2_{0\beta}$ ja R^2_H, joihin ei ole asetettu kovariaatteja. Jos lasketaan $R^2_{0\beta}$ ja R^2_H taulukosta, johdetaan muut matriisin elementit laajennetusta taulukosta. Matriisin rakenteesta käy ilmi, että jäljelle jäävät diagonaalielementit saadaan ANCOVA-taulusta, missä $\boldsymbol{y}$ korvataan matriisin $\boldsymbol{Z}$ sarakkeilla. Kovarianssi analyysi voidaan havainnollistaa ANCOVA-taulun avulla, jota hyödynnetään myöhemmin testauksessa.

Hyödynnetään mallia 2.8, saaden kaikki neliösummat ANCOVA-tauluun. Mallin oikeanpuoleiset termit jakautuvat vielä kahteen komponenttiin, joita merkitään

$$SS\beta = R^2_{H\beta}, \qquad SS\eta = \boldsymbol{r}'_H\boldsymbol{R}'_H\boldsymbol{r}_H. \tag{2.10}$$

Komponenteista 2.10 saadaan ANCOVA-tauluun residuaalien neliösummat ja näiden kaikkien neliösummien yhteistulos, joita merkitään

$$SSE = SST - SS\beta - SS\eta, \qquad SST = SSE + SS\beta + SS\eta.$$

Neliösummia hyödyntäen saadaan näiden kaikkien keskineliöt jakamalla neliösummat niiden vapausasteilla. Keskineliövirheitä merkitään

$$MS\beta = \frac{SS\beta}{\rho(\boldsymbol{X})}, \qquad MS\eta = \frac{SS\eta}{\rho(\boldsymbol{Z})}, \qquad MSE = \frac{SSE}{n - \rho(\boldsymbol{X}) - \rho(\boldsymbol{Z})}$$

Kaavassa 2.9 esitellään F-testisuureen kaava ja hyödynnetään tätä ANCOVA-taulun merkitsemisessä. F-testisuuren arvoja merkitään

$$F_\beta = \frac{MS\beta}{MSE}, \qquad F_\eta = \frac{MS\eta}{MSE}$$

Taulukko 2.1. Kovarianssianalyysin vapausasteet, neliösummat, neliösummat

	Vapausasteet	Neliösumma	Keskineliö	F-testisuure
Luokat	$\rho(\boldsymbol{X})$	SSβ	MSβ	F_β
Yhteisvaikutus	$\rho(\boldsymbol{Z})$	SSη	MSη	F_η
Residuaalit	$n - \rho(\boldsymbol{X}) - \rho(\boldsymbol{Z})$	SSE	MSE	
Kaikki	n	SST		

Taulukossa 2.1 on esitelty kovarianssianalyysissa tärkeimmät tunnusluvut. Hyödynnetään näitä luvussa 3 tutkimaan kovariaattien vaikutusta testauksessa ja tulosten raportoinnissa.

3 Palkkatasoaineiston testaus ja analysointi

Tässä luvussa esitellään aineisto ja sen muuttujat alaluvussa 3.1, joita hyödynnetään testauksessa. Hyödynnetään aineistoa kovarianssianalyysin testaukseen. Aineiston analysoinnissa käytetyt mallit niiden testaukset esitellään alakohdassa 3.2. Malleja joita hyödynnetään, on kaksi erilaista havainnollistamaan kovarianssianalyysin testausta ja vähentämällä kovariaattien vaikutusta.

3.1 Palkkatasoaineiston kuvaus

Tässä tutkimuksessa hyödynnetään Suomen Ekonomit järjestön teettämää palkkatasotutkimusta havainnollistamaan kovarianssianalyysin testausta. Palkkatasotutkimus on kerätty vuonna 2018 Suomen Ekonomit järjestön jäseniltä, jotka ovat suorittaneet kauppatieteellisen yliopistotutkinnon. Kyselyn tarkasteluajankohta on lokakuu 2018. Aineistossa on 3465 havaintoa ja 44 erilaista muuttujaa.

Aineistossa on puuttuvia havaintoja, koska aineiston keruu on toteutettu kyselylomakkeella. Puuttuvat havainnot voivat myös johtua siitä, että lomakkeessa on kysymyksiä, joissa tietyn vastauksen annettua tarvitsee vastata toiseen kysymykseen. Kyselylomakkeessa on jatkuvia ja diskreettejä muuttujia sekä avoimia tekstiosioita.

3.2 Muuttujien valinta

Tutkielmassa tarkastellaan kovariaatteina iän ja henkilöstömäärän vaikutusta palkkaeroihin. Tutkielmassa ollaan kiinnostuneita sukupuolen välisistä palkkaeroista. Epäillään, että ikäerot voivat vaikuttaa keskimääräisesti palkkatasoon sukupuolten välillä. Tätä muuttujaa käytetään ensimmäisessä ANCOVA-mallissa. Kovarianssianalyysissa voidaan ottaa huomioon ikämuuttujan vaikutus lisäämällä se kovariaattina analyysiin.

Käytetään toisessa ANCOVA-mallissa kategorista muuttujaa työnantajan henkilöstömäärä. Tämä muuttuja lisätään ANCOVA-malliin kovariaattina. Tutkielmassa tutkitaan vaikuttaako sukupuoli tilastollisesti merkitsevästi keskimääräiseen palkkatasoon silloin, kuin miesten ja naisten keski-iän erot ja henkilöstömäärän ovat otettu huomioon.

Taulukko 3.1. Muuttujien tunnuslukuja.

Muuttuja	Minimi	Mediaani	Keskiarvo	Maksimi	Puuttuvia arvoja
Palkka	1000	5000	5800	65000	6
Nainen	1000	4700	5137	26323	2
Mies	1000	5550	6644	65000	4
Ikä	21	42	42.40	75	60
Nainen	21	42	42.41	64	21
Mies	24	41	42.42	75	32

Taulukosta 3.1 huomataan, että suurin palkka-arvo on hyvin suuri verrattuna mediaaniin tai keskiarvoon. Tämä arvo voi mahdollisesti vaikuttaa testituloksiin. Huomioidaan tämä arvo testauksien analysoinnissa.

Naiset tienaavat keskimäärin vähemmän palkkaa kuin miehet. Naisten keskimääräinen kuukausipalkka on 5137, kun taas miehillä keskimääräinen kuukausipalkka on 6644. Huomataan, että suurin kuukausipalkka 65000 löytyy aineistosta mieheltä. Naisilla maksimipalkka on huomattavasti pienempi, kuin miehillä.

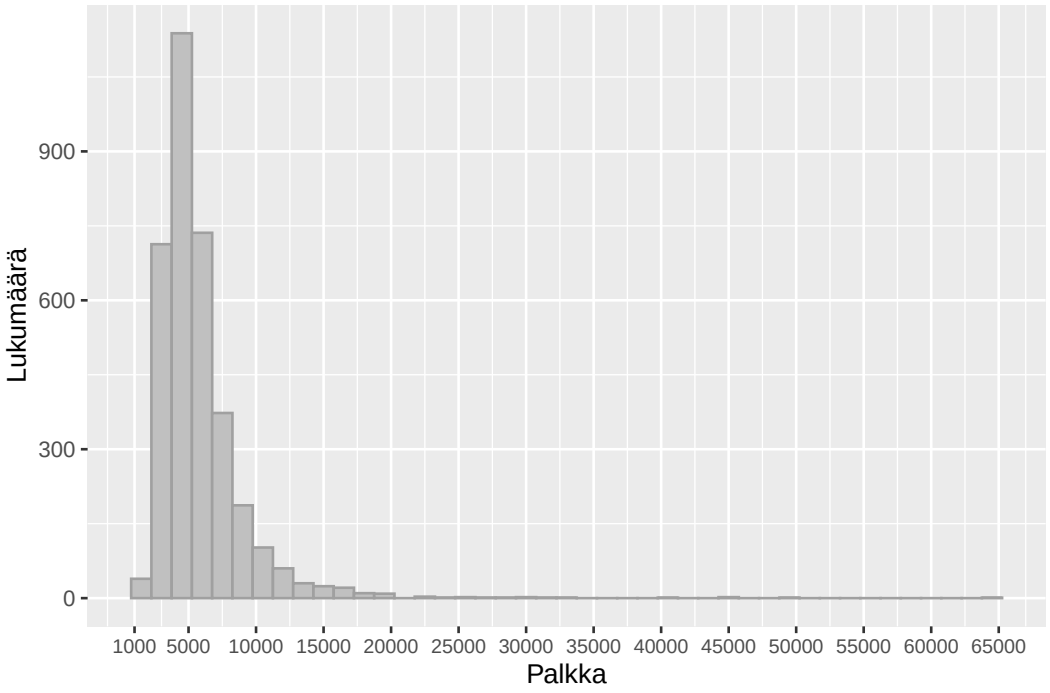
Naisten ja miesten iät ovat jakautuneet tasaisesti. Aineistossa naisten maksimi- ja minimi-ikä on pienempi kuin miehillä, sekä keskimääräinen ikä on lähes sama.

Taulukko 3.2. Muuttujien tunnuslukuja.

Muuttuja	Lukumäärä	Puuttuvia arvoja
Sukupuoli	3445	20
Nainen	1983	
Mies	1462	
Henkilöstömäärä	3444	21
1-9	227	
10-29	291	
30-99	453	
100-249	444	
250-499	383	
500-999	446	
1000-2999	455	
yli 3000	745	

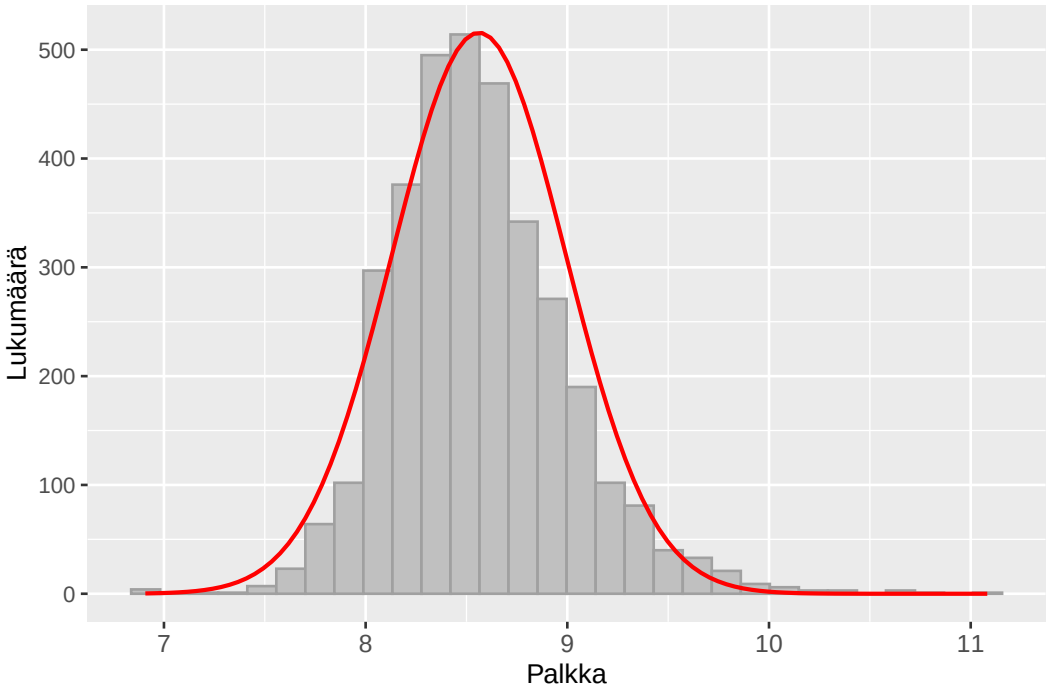
Taulukosta 3.2 huomataan aineistossa olevan miehiä ja naisia lähes yhtä monta. Naisia on silti enemmän kuin miehiä. Henkilöstömäärä muuttuja kuvaa kuinka suuri on kyselynvastaajan työnantajan henkilöstömäärä. Lukumäärät ovat jakautuneet henkilöstömäärien luokissa hyvin tasaisesti lukuun ottamatta luokkaa yli 3000, missä lukumäärä on suurin verrattuna muihin luokkiin.

Tarkastellaan ensimmäiseksi palkan jakautumista aineistossa histogrammin avulla.



Kuva 3.1. Palkan jakautuminen histogrammilla kuvattuna.

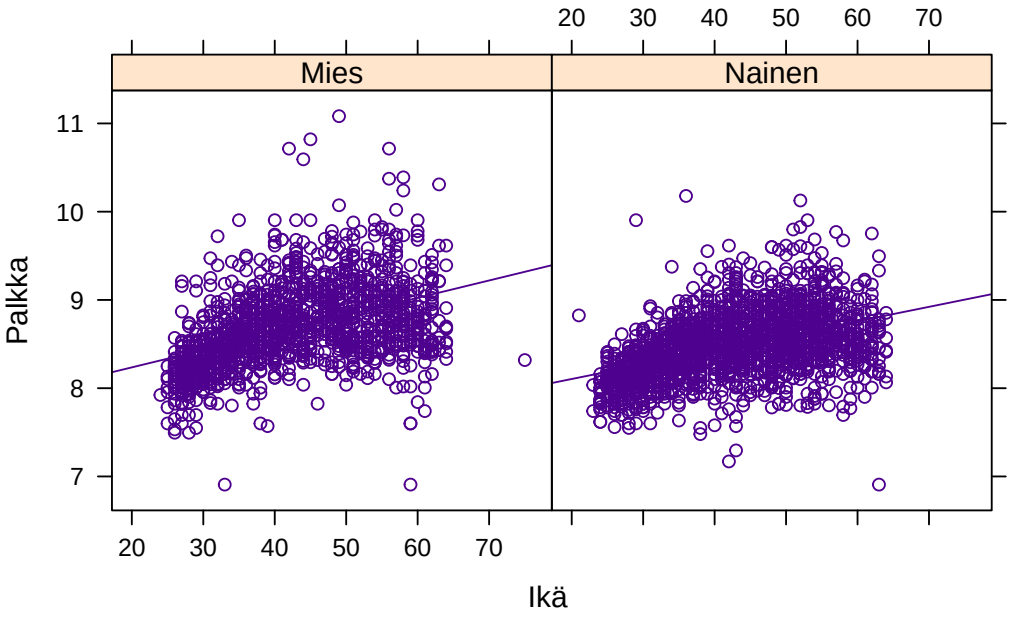
Kuvasta 3.1 voidaan huomata että palkka ei jakaudu normaalisti, joten on syytä laskea luonnollinen logaritmi palkkamuuttujalle. Huomataan myös poikkeavan palkka-arvon takia, joka nähtiin taulukosta 3.1, histogrammi kaartuu loivasti oikealle. Seuraavaksi esitetään palkan jakautuminen, kun olemme ottaneet logaritmin palkka-arvoista.



Kuva 3.2. Logaritmoidun kuukausipalkka-muuttujan jakautuminen histogrammilla kuvattuna.

Palkkamuuttuja on nyt jakautunut normaalisti, kuten voidaan kuvasta 3.2 huomata. Täten luonnollisen logaritmin ottaminen palkkamuuttujasta on mielekäs vaihtoehto, joten käytetään uutta muuttujaa aineiston analysoinnissa.

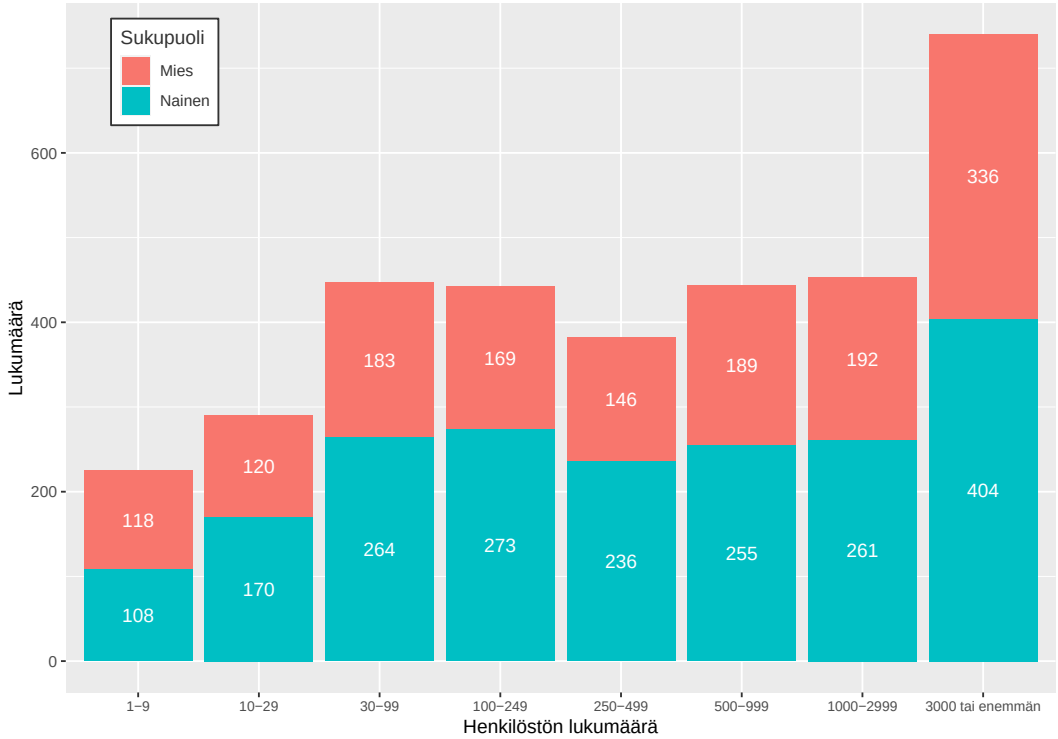
Tarkastellan seuraavaksi palkan jakautumista iän suhteen graafin avulla.



Kuva 3.3. Logaritmoidun kuukausipalkan ja iän jakautuminen sukupuolten välillä

Kuvassa 3.3 on käytetty logaritmisoituja palkka-arvoja. Voidaan huomata, että sukupuolen välillä on eroavaisuuksia. Keskimääräisesti miesten palkat ovat korkeammat, kuin naisten. Naisten palkka suhteessa ikään on hajautuneet tiheämmälle alueelle kuin miesten. Yleisesti palkat kasvavat, kun ikää kertyy, mutta havaitaan pisteiden muodostavan paraabelimaisen kuvion miehillä ja naisilla. Miehillä palkka on yleisemin korkeampaa 40 – 50 -vuotiailla ja on alemmillaan nuoremmilla henkilöillä sekä alentuu, kun ikää kasvaa. Naisilla paraabelimaisuus ei ole niin voimakas, kuin miehillä.

Tarkastellaan seuraavaksi histogrammin avulla sukupuolen jakautumista työnantajan henkilöstömäärällä.



Kuva 3.4. Sukupuolten jakautuminen työnantajan eri henkilöstömäärissä

Työnantajan henkilöstömäärä on kuvassa 3.4 tasaista melko 30 – 2999 henkilöstömäärällä, mutta vähintään 3000 työntekijän henkilöstömäärä on suurin lukumäärältään. Tämä todettiin jo taulukossa 3.2. Jokaisella pylväällä naisten ja miesten väliset lukumäärät on esitetty värein ja luvuin. Naisten ja miesten välillä henkilöstön koko on jakautunut melko tasaisesti muuttujan eri luokilla. Ainoastaan miesten osuus on suurempi 1 – 9 kokoisessa organisaatiossa ja naisten osuus on suurempaa muissa luokissa.

3.2.1 Sukupuolen ja iän vaikutus palkkaan

Olemme esitelleet palkan ja iän jakautumisen sukupuolille luvussa 3.2. Voidaan seuraavaksi hyödyntää kovarianssianalyysia muuttujien testaukseen.

Ensimmäiseksi sovitetaan malliin selitettäväksi muuttujaksi palkka ja selittäväksi muuttujaksi sukupuoli. Ikä muuttuja on tässä mallissa kovariaatti. Mallia kuvataan seuraavalla esityksellä.

$$\text{Palkka} = \beta_0 + \beta_1 \text{Sukupuoli} + \beta_2 \text{Ikä} + \epsilon$$

Estimoitu kovarianssianalyysin malli saadaan taulukon 3.3 estimaattien arvoista. Nämä arvot kuvaavat muuttujien kertoimia.

Taulukko 3.3. ANCOVA-mallin parametriestimaatit

	Estimaatti
Vakio	7.9235
sukupuoli	-0.2067
ikä	0.0178

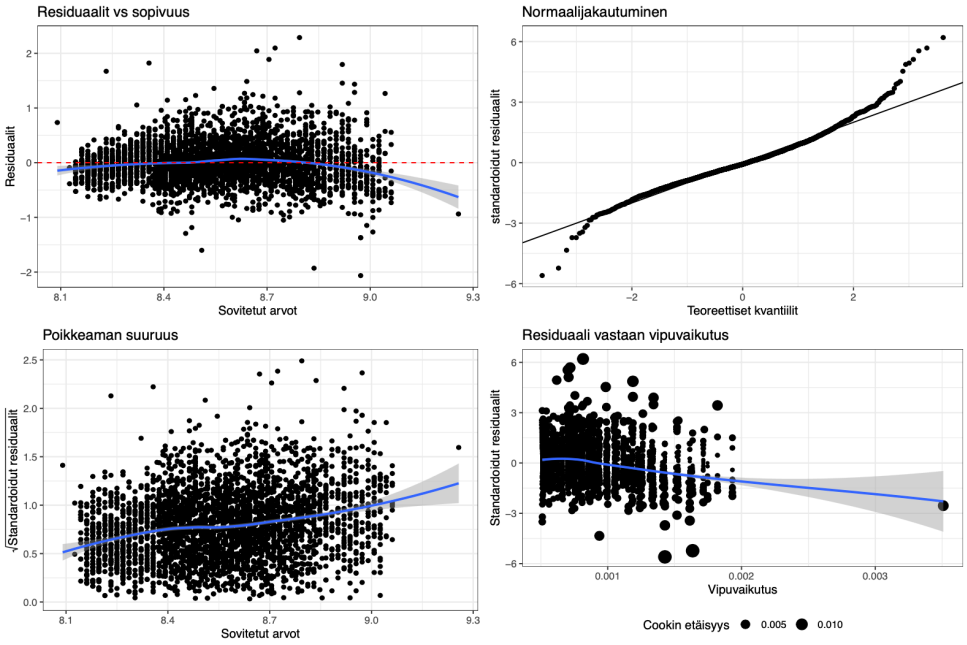
Taulukosta 3.3 saadaan vakiotermin $\widehat{\beta_0}$ kerroin, sukupuolen kerroin $\widehat{\beta_1}$ ja iän kerroin $\widehat{\beta_2}$. Vakiotermin $\widehat{\beta_0}$ kerroin on logaritmisoitu ja tämä asia on hyvää pitää mielessä, kun tehdään kovarianssianalyysin testaus. Sukupuolen estimaatti kertoo, että naiset tienaavat keskimäärin huonompaa palkka -0.207 yksikköä miehiin verrattuna, kun palkka on normaalisti jakautunut. Estimoitua kovarianssianalyysin mallia voidaan kuvata

$$\widehat{y} = 7.9235 - 0.2067\text{sukupuoli} + 0.0178\text{ikä}$$

Taulukko 3.4. ANCOVA -taulu

	df	Neliösumma	Keskineliösumma	F-testisuure	p-arvo
Sukupuoli	1	35.55	35.552	261.08	$2 \cdot 10^{-16}$
ikä	1	119.30	119.305	876.10	$2 \cdot 10^{-16}$
Residuaalit	3383	460.68	0.136		

Taulukossa 3.4 on esitetty mallin kovarianssianalyysin testauksen tulokset. Sukupuolen ja iän F-testisuureet ovat suuria ja p-arvot ovat pieniä, joten voidaan todeta palkoissa olevan eroavaisuuksia, kun iän vaikutus on poistettu. Tutkitaan visuaalisesti palkan jakautumista.



Kuva 3.5. ANCOVA mallin residuaaliesitys

Kuvassa 3.5 vasemmassa ylälaidassa oleva kuva näyttää mallin sopivuuden. Tässä tapauksessa residuaalit eivät leviä tasaisesti vaakasuoran viivan ympärille, jolloin malli on epälineaarinen. Voimme huomata, että poikkeavia arvoja löytyy, jolloin vaakasuora viiva kaartuu alemmas.

Oikealla yläkulmassa kuvassa nähdään, jakautuuko residuaalit normaalista. Kuvasta voidaan päätellä, että residuaalit eivät jakaudu tasaisesti, koska standardoitujen residuaalien arvot ylä- ja alapäässä kaartuvat.

Vasemman alakulman kuva on hyvin samanlainen, kuin ensimmäinen kuva mutta nyt tämä kuva näyttää leviävätkö residuaalit tasapuolisesti ennustajien vaihteluvälille. Tämän kuvan avulla voidaan tarkistaa oletukset varianssien symmetrisyydestä. Kuvassa residuaalit ovat jakautuneet melko satunnaisesti ennustealueelle.

Oikealla alakulmassa olevassa kuvassa tarkastellaan, onko yksittäisellä pisteellä vaikutusta ANCOVA-malliin. Huomataan suuremmista pisteistä, että vaikutusta löytyy. Jos poistaisimme arvon, se vaikuttaisi tuloksiin. Seuraavaksi tarkastellaan eri kovariaattia kuin iän vaikutusta palkkaeroihin.

3.2.2 Henkilöstömäärän vaikutus palkkaeroihin

Aikaisemmin todettiin palkkaeroihin vaikuttavan sukupuoli, kun otamme huomioon myös iän. Seuraavaksi testataan palkkaeroja samalla minimoiden sekoittavien tekijöiden vaikutus pois yrityksen henkilöstömäärän avulla. Käytetään samoja muuttujia palkkaa ja sukupuolta uuteen ANCOVA-malliin, mutta lisätään malliin lisäksi henkilöstömäärä vähentämään palkkaeroja. Uutta mallia kuvataan seuraavalla esityksellä.

$$\text{Palkka} = \beta_0 + \beta_1 \text{Sukupuoli} + \beta_{2i} \text{Henkilöstömäärä} + \epsilon,$$

missä $i = 1, \ldots, 8$. Indeksin i arvot kuvaavat henkilöstömäärän kokoa. Tarkastellaan mallin estimaatteja. Taulukossa 3.5 on esitetty nämä estimaatit.

Taulukko 3.5. ANCOVA-mallin parametriestimaatit ja p-arvot

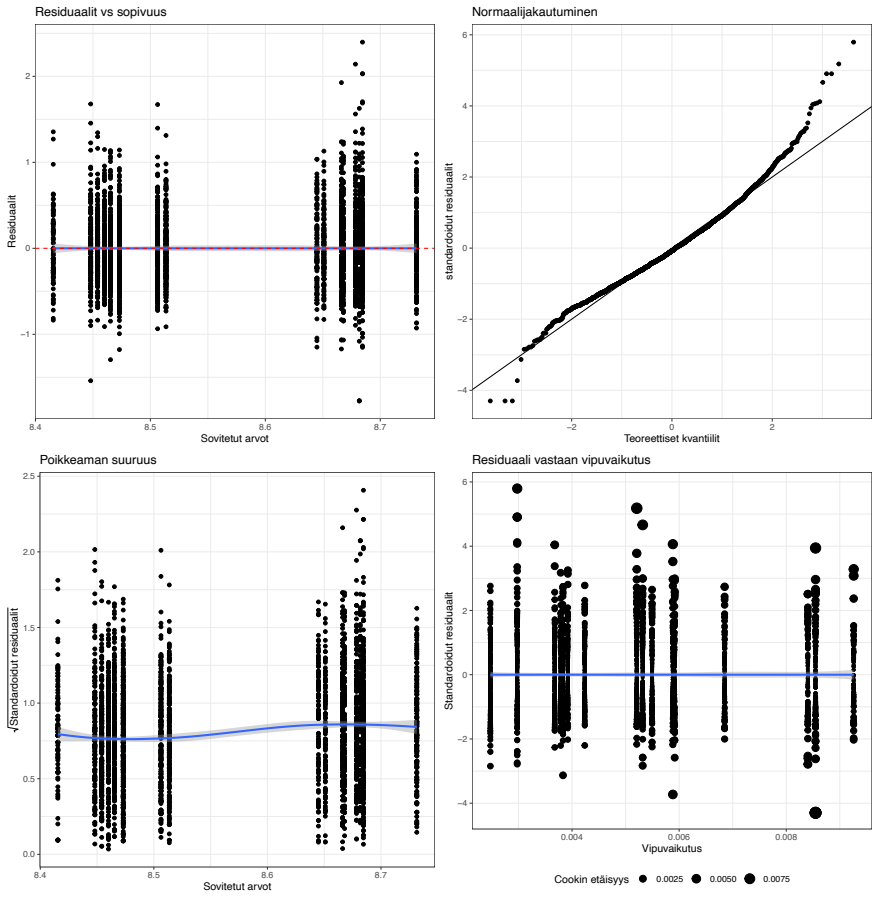
	Estimaatti	p-arvo
Vakio	8.681783	$2 \cdot 10^{-16}$
Sukupuoli	-0.266287	$2 \cdot 10^{-16}$
Henkilöstömäärä		
1-9		
10-29	-0.002483	0.9463
30-99	0.071702	0.0345
100-249	0.042548	0.2107
250-499	0.007550	0.8286
500-999	0.009922	0.7700
1000-2999	0.021526	0.5243
3000 tai enemmän	0.028748	0.3622

Taulukossa 3.5 vakiotermin $\widehat{\beta_0}$ kerroin on 8.681, sukupuolimuuttujan kerroin $\widehat{\beta_1}$ on -0.266 sekä henkilöstömäärämuuttujan kertoimet ovat $\widehat{\beta_{21}}$, $\widehat{\beta_{22}}$, $\widehat{\beta_{23}}$, $\widehat{\beta_{24}}$, $\widehat{\beta_{25}}$, $\widehat{\beta_{26}}$, $\widehat{\beta_{27}}$ ja $\widehat{\beta_{28}}$ jokaiselle henkilöstömäärän luokalle.

Taulukko 3.6. ANCOVA -taulu

	df	Neliösumma	Keskineliösumma	F-arvo	p-arvo
Sukupuoli	1	35.34	35.342	206.0468	$2 \cdot 10^{-16}$
Henkilöstömäärä	7	1.70	0.242	1.4119	0.195
Residuaalit	3410	584.89	0.172		

Taulukosta 3.6 on esitetty tunnusluvut mallille. Voidaan huomata, kun henkilöstömäärän eli kovariaatin vaikutus on poistettu, palkoissa on sukupuolittain eroavaisuuksia. Havaitaan, että kovariaatilla ei ole vaikutusta palkkaeroihin, koska p-arvo on $0.195 > 0.05$. Tutkitaan mallin residuaaliesitystä.



Kuva 3.6. ANCOVA mallin residuaaliesitys

Kuvassa 3.6 vasemmalla yläkulmassa olevan kuvan residuaalit ovat jakautuneet melko tasaisesti. Poikkeavia arvoja ANCOVA -mallissa ei ole kovin paljoa lukuun ottamatta viivan häntäpäitä, jossa sininen alue leviää hieman.

Oikealla yläkulmassa kuvan residuaalit eivät ole jakautuneet kovin normaalisti, koska standardoidut residuaalit arvot kaartuvat kuvan ylä- ja alapäässä. Huomataan myös häntäpäässä olevan kaarre ennen kuin arvot kaartuvat alaspäin.

Vasemmalla alakulmassa residuaalit leviävät tasaisesti ennustajien vaihteluvälille. Sininen viiva on melko vaakasuora lukuun ottamatta pientä kaarevuutta.

Oikealla alakulmassa voidaan nähdä, ettei yksittäisellä pisteellä ole vaikutusta. Todetaan näiden kuvien perusteella mallin olevan melko sopiva.

3.2.3 Iän ja Henkilöstömäärän vaikutus palkkaeroihin

Ollaan tarkastelu aikaisemmissa erikseen ikämuuttujan ja henkilöstömäärämuuttujan vaikutus kovariaattina ja niiden merkitystä mallissa. Tehdään vielä kolmas malli, missä käytetään kumpaakin kovariaattia ikämuuttujaa sekä henkilöstömäärämuuttujaa yhtä aikaa mallissa. Tarkastellaan miten sukupuolierot muuttuva, kun lisätään kovariaatteja malliin. Tarkastellaan ensimmäiseksi mallia, jota kuvataan seuraavalla esityksellä

$$\text{Palkka} = \beta_0 + \beta_1 \text{Sukupuoli} + \beta_2 \text{Ikä} + \beta_{3i} \text{Henkilöstömäärä} + \epsilon,$$

missä $i = 1, \ldots, 8$. Indeksin i arvot kuvaavat henkilöstömäärän kokoa. Tarkastellaan mallin estimaatteja ja taulukossa 3.7 on esitetty nämä estimaattien arvot ja niiden p-arvot.

Taulukko 3.7. ANCOVA-mallin parametriestimaatit ja p-arvot

	Estimaatti	p-arvo
Vakio	8.080055	$2 \cdot 10^{-16}$
Sukupuoli	-0.206465	$2 \cdot 10^{-16}$
Ikä	0.017800	$2 \cdot 10^{-16}$
Henkilöstömäärä		
1-9		
10-29	0.028625	0.3874
30-99	0.078192	0.0102
100-249	0.062023	0.0425
250-499	0.022277	0.4771
500-999	0.032164	0.2925
1000-2999	0.057298	0.0595
3000 tai enemmän	0.068017	0.0166

Taulukossa 3.7 on esitetty mallin estimoitujen termien kertoimet. Vakiotermin $\widehat{\beta_0}$ kerroin on 8.08, sukupuolimuuttujan kerroin $\widehat{\beta_1}$ on -0.206 sekä kovariaattien ikämuuttujan $\widehat{\beta_2}$ kerroin on 0.018 ja henkilöstömäärämuuttujan $\widehat{\beta_{3i}}$ kertoimet ovat esitetty luokittain, missä indikaattori $i = 1, \ldots, 8$ ja indikaattorin i arvot kuvaavat henkilöstömäärien luokkia.

Taulukko 3.8. ANCOVA -taulu

	df	Neliösumma	Keskineliösumma	F-arvo	p-arvo
Sukupuoli	1	34.76	34.75	255.54	$2 \cdot 10^{-16}$
Ikä	1	118.05	118.05	867.92	$2 \cdot 10^{-16}$
Henkilöstömäärä	7	1.78	0.255	1.8719	0.07001
Residuaalit	3357	456.60	0.14		

Taulukosta 3.8 on esitetty kovarianssianalyysin testauksen tulokset. Voidaan huomata, kun kovariaattien iän ja henkilöstömäärän vaikutus on poistettu, palkoissa on eroavaisuuksia sukupuolittain. Henkilöstömäärän p-arvo on suurempi, kun testataan ilman ikämuuttujaa. Ikämuuttujan p-arvo ei ole lainkaan muuttunut, kun tarkastellaan malleja, missä muuttuja on kovariaattina yksinään. Havaitaan, että henkilöstömäärämuuttujalla ei ole vaikutusta palkkaeroihin, koska p-arvo on $0.07 > 0.05$. Ikämuuttujalla ei ole myöskään tässä mallissa merkitystä.

Havaitaan kaikkien mallien sukupuolen estimaattien olevan hyvin samat. Ensimmäisen mallin estimaatti on -0.207, toisessa mallissa se on -0.266 ja kolmannessa mallissa se on -0.206. Tästä voidaan päätellä, että kovariaattien lisääminen ei vaikuta juuri ollenkaan miesten ja naisten palkkoihin.

4 Johtopäätelmät

Tutkielmassa tutkittiin naisten ja miesten välisiä palkkaeroja, kun taustamuuttujien vaikutus on vakioitu. Taulukossa 3.1 havaittiin miesten tienaavan keskimääräisesti enemmän kuin naiset. Haluttiin selvittää, johtuuko palkkaero taustamuuttujista eli sekoittavista tekijöistä. Käytettiin havainnollistamaan kolmea erilaista mallia. Ensimmäisessä mallissa esiteltiin sukupuolen vaikutus palkkaan lisäämällä kovariaatiksi ikämuuttujan. Havaittiin miesten palkkojen olevan yhä suurempia kuin naisten, kun iän vaikutus on vakioitu.

Toisen mallin tapauksessa tutkittiin palkkaeroja lisäämällä sekoittavaksi tekijäksi henkilöstömäärä. Kun poistetaan vaikutus, minkä kokoisessa yrityksessä kyselyn vastaaja työskentelee, se ei vaikuta palkkaeroihin merkitsevästi. Henkilöstömäärän vaikutuksen poistamisen jälkeen miesten palkka on yhä suurempi.

Kolmanteen malliin sisällytettiin molemmat sekoittavat tekijät ikä ja henkilöstömäärä. Tulokset osoittivat, että kovariaattien vaikutus on hyvin pieni palkkaeroihin, kun poistettiin iän ja henkilöstömäärän vaikutus. Kovariaattien lisääminen malliin ei siis juurikaan vähentänyt palkkaeroja. Miesten ja naisten palkoissa on eroa, mutta palkkojen eroa ei selitä henkilöstömäärä ja ikä. Palkkaeroon vaikuttavat muut tekijät enemmän.

Lähteet

Huitema, B. (2011). *The analysis of covariance and alternatives : Statistical methods for experiments, quasi-experiments, and single-case studies.* URL: https://ebookcentral.proquest.com.

Sengupta, D. ja S. Jammalamadaka (2003). *Linear Models : an Integrated Approach*. River Edge, N.J: World Scientific.

Venables, W. ja B. Ripley (1999). *Modern applied statistics with S-PLUS*. 3. ed. New York: Springer.

Liite: R-koodi

```
> l<-lm(log(palkka)~factor(sukup)+ika, data=ekodat)
> anova(l)

Analysis of Variance Table

Response: log(palkka)
                Df Sum Sq Mean Sq F value    Pr(>F)
factor(sukup)    1  35.55  35.552  261.08 < 2.2e-16 ***
ika              1 119.30 119.305  876.10 < 2.2e-16 ***
Residuals     3383 460.68   0.136
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(l)

Call:
lm(formula = log(palkka) ~ factor(sukup) + ika, data = ekodat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06439 -0.22283 -0.02859  0.20158  2.28774

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.9234661  0.0272947  290.29   <2e-16 ***
factor(sukup)2 -0.2066892  0.0128442  -16.09   <2e-16 ***
ika             0.0177742  0.0006005   29.60   <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.369 on 3383 degrees of freedom
  (79 observations deleted due to missingness)
Multiple R-squared:  0.2516,    Adjusted R-squared:  0.2511
F-statistic: 568.6 on 2 and 3383 DF,  p-value: < 2.2e-16

> l2<-lm(log(palkka)~ sukup+factor(thenkmaar), data=ekodat)
> summary(l2)

Call:
lm(formula = log(palkka) ~ sukup + factor(thenkmaar), data = ekodat)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.7454 -0.2808 -0.0326  0.2382  2.4003

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.859759   0.034852 254.213   <2e-16 ***
sukup              -0.206617   0.014370 -14.378   <2e-16 ***
factor(thenkmaar)2 -0.002483   0.036854  -0.067   0.9463
factor(thenkmaar)3  0.071702   0.033904   2.115   0.0345 *
factor(thenkmaar)4  0.042548   0.033987   1.252   0.2107
factor(thenkmaar)5  0.007550   0.034860   0.217   0.8286
factor(thenkmaar)6  0.009922   0.033932   0.292   0.7700
factor(thenkmaar)7  0.021526   0.033806   0.637   0.5243
factor(thenkmaar)8  0.028748   0.031544   0.911   0.3622
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4142 on 3410 degrees of freedom
  (46 observations deleted due to missingness)
Multiple R-squared:  0.05955,   Adjusted R-squared:  0.05735
F-statistic: 26.99 on 8 and 3410 DF,  p-value: < 2.2e-16

> anova(l2)
Analysis of Variance Table

Response: log(palkka)
                    Df Sum Sq Mean Sq  F value Pr(>F)
sukup                1  35.34  35.342 206.0468 <2e-16 ***
factor(thenkmaar)    7   1.70   0.242   1.4137  0.195
Residuals         3410 584.89   0.172
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> l3 <- lm(log(palkka)~ sukup+ika+factor(thenkmaar), data=ekodat)
> summary(l3)

Call:
lm(formula = log(palkka) ~ sukup + ika + factor(thenkmaar), data = ekodat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0160 -0.2275 -0.0291  0.2021  2.2683

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.080055   0.041284 195.718   <2e-16 ***
sukup              -0.206465   0.012903 -16.001   <2e-16 ***
```

```
ika                 0.017800   0.000604  29.471   <2e-16 ***
factor(thenkmaar)2  0.028625   0.033116   0.864   0.3874
factor(thenkmaar)3  0.078192   0.030408   2.571   0.0102 *
factor(thenkmaar)4  0.062023   0.030564   2.029   0.0425 *
factor(thenkmaar)5  0.022277   0.031334   0.711   0.4771
factor(thenkmaar)6  0.032164   0.030553   1.053   0.2925
factor(thenkmaar)7  0.057298   0.030399   1.885   0.0595 .
factor(thenkmaar)8  0.068017   0.028370   2.397   0.0166 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3688 on 3357 degrees of freedom
  (98 observations deleted due to missingness)
Multiple R-squared:  0.2529,    Adjusted R-squared:  0.2509
F-statistic: 126.3 on 9 and 3357 DF,  p-value: < 2.2e-16

> anova(l3)
Analysis of Variance Table

Response: log(palkka)
                    Df Sum Sq Mean Sq  F value  Pr(>F)
sukup                1  34.76  34.757 255.5379 < 2e-16 ***
ika                  1 118.05 118.050 867.9222 < 2e-16 ***
factor(thenkmaar)    7   1.78   0.255   1.8719 0.07001 .
Residuals         3357 456.60   0.136
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```