3.3 Empiirisiä esimerkkejä

3.3.1 Itsemurhat

Daly ym. (2011)[14] estimoivat PNS-menetelmällä yhtälön

$$\begin{array}{rcccccc} y & = & 24,912 & + & 8,255x & + & \hat{\varepsilon}, \\ & & (2,311) & & (3,992) & & \end{array}$$

$$\mathsf{R}^2 = 0,248,\ n = 15.$$

Yllä y on itsemurhien lukumäärä 100 000 kansalaista kohti, x on kansakunnan onnellisuutta mittaava indeksi, $\hat{\varepsilon}$ on mallin residuaali, luvut suluissa ovat keskivirheitä ja n on havaintojen lukumäärä. Jäännösten ε oletetaan noudattavan normaalijakaumaa $\mathsf{N}(0, \sigma^2)$ ja olevan keskenään korreloimattomia. Kukin havaintopari (x_i, y_i) liittyy eurooppalaiseen valtioon ($i = 1, \ldots, 15$). Havainnot ja niihin sovitettu regressiosuora ovat oheisessa kuviossa.

Mallin mukaan

- itsemurhaintensiteetti (itsemurhien lukumäärä 100 000 kansalaista kohti) on 24, 912 (estimoitu vakio), kun onnellisuusindeksi saa arvon 0.
- itsemurhaintensiteetti kasvaa onnellisuusindeksin kasvaessa. Kun jälkimmäinen suurenee yksiköllä, edellinen kasvaa 8, 255:llä (estimoitu kerroin onnellisuusindeksille).
- 24, 8 prosenttia itsemurhaintensiteetin vaihtelusta selittyy onnellisuusindeksin vaihtelulla (selitysasteen R^2:n suuruus).

Onnellisuusindeksi saa toiseksi korkeimman arvonsa aineistossa Suomen kohdalla. Suomessa tehdään itsemurhia vielä enemmän kuin malli ennustaa — eniten koko aineistossa.

Kuviossa raportoidun korrelaatiokertoimen $0,4975\dot{}$:n neliö on mallin selitysaste 0, 248, koska mallissa on vain yksi selittäjä (kaava (5)).

Koska jäännökset ε ovat normaalijakautuneita ja keskenään korreloimattomia, estimoidut kertoimet jaettuna keskivirheillään ovat t-jakautuneita. Koska mallissa on vain yksi selittäjä ja havaintoja on 15, on jakauma t_{15-1-1} eli t_{13} (kaava (6)). Testisuure on $8,255/3,992 \approx 2,068$. Jakauman t_{13} 97, 5. persentiili on 2, 160 (oheisesta taulukosta). Koska $|2,068| < |2,160|$, niin nollahypoteesi ei tule aivan hylätyksi 5 %:n riskitasolla kaksisuuntaisessa testauksessa. Onnellisuusindeksi ei ole tilastollisesti merkitsevä selittäjä eikä aineiston perusteella ole syytä luopua oletuksesta, että itsemurhaintensiteetti ja onnellisuusindeksi eivät korreloi.

Kuvion perusteella saattaisi veikata, että muuttujien välillä olisi todellinen yhteys. Selitys tilastolliselle merkitsettömyydelle saattaa olla aineiston pieni

[14] Mary C. Daly, Andrew J. Oswald, Daniel Wilson ja Stephen Wu (2011): Dark Contrasts: The Paradox of High Rates of Suicide in Happy Places. *Journal of Economic Behavior & Organization*, 80, 435–442.

koko: Tilastollisesti merkitsevä positiivinen suhde itsemurhaintensiteetin ja osavaltioiden onnellisuusindeksien välillä pätee Yhdysvalloissa (mt.), ja osavaltioita on enemmän kuin eurooppalaisia valtioita regressiossa edellä. Mahdollisesti osavaltiot ovat myös homogeenisempia kuin eurooppalaiset valtiot, jolloin tutkittu suhde tulee selvemmin esiin osavaltioaineistossa (jäännös sisältää vähemmän vaihtelevia tekijöitä).

3.3.2 Siivoojien tuntipalkat

Keinänen ja Pakarinen (2009) tutkivat siivoojien tuntipalkkoja ja mahdollista palkkasyrjintää suomalaisessa siivousyrityksessä vuonna 2007.[15] He estimoivat vaihtoehtoisia malleja, jotka ovat kaikki palkkasyrjintää koskevalta tulokseltaan yhtäpitäviä. Yksi heidän (PNS-menetelmällä) estimoimistaan malleista on

$$\begin{array}{rcccccccccc} y & = & 8,430 & + & 0,114x_1 & - & 0,001x_2 & + & 0,169x_3 & + & 0,339x_4 & + & \hat{\varepsilon}. \\ & & (0,000) & & (0,840) & & (0,983) & & (0,747) & & (0,000) & & \end{array} \tag{12}$$

$$\mathsf{R}^2 = 0,256,\ F_{4,132} = 11,269,\ n = 137.$$

Yhtälössä y on tuntipalkka, x_1 on indeksi, joka saa arvon 1, kun siivooja on mies ja 0 muuten, x_2 on siivoojan ikä, x_3 on indikaattori työsuhteen laadulle, joka saa arvon 1, kun työsuhde on toistaiseksi voimassa oleva ja 0 muutoin[16] ja x_4 on työsuhteen kesto vuosina. Muut merkinnät (F-tunnusluvulla täydennettynä) ja oletukset ovat kuten edellisessä esimerkissä. Kukin havaintovektori $[x_{i1}\ \ldots\ x_{i4}\ y_1]$ liittyy yhteen siivojaan ($i = 1, \ldots, 137$).

Jäännöksen normaalisuusoletuksen perusteella testisuureet noudattavat t- ja F-jakaumia. Muuttujien selityskykyä yhdessä testaan F-testisuurella $F = 11,269$. Nollahypoteesin pätiessä se noudattaa jakaumaa $F_{4,132}$ (kaava (9)). Tätä jakaumaa ei ole taulukoitu oheisessa taulukossa, mutta jakauma $F_{4,100}$ on. Käytetään sitä vertailujakaumana. Sen 95. persentiili on $2,463$.[17] Koska $11,269 > 2,463$, niin nollahypoteesi hylätään. Mallin selittäjillä on yhdessä selityskykyä.

Sukupuoli-indikaattorin t-arvo on $0,114/0,840 \approx 0,136$. Nollahypoteesin (kerroin on 0) pätiessä se noudattaa $t_{137-4-1}$ eli t_{132}-jakaumaa (kaava (10)). Oheisessa taulukossa ei ole taulukoitu t-jakaumaa 132 vapausasteella. Valitaan lähinnä taulukoitu vapausasteluku, joka on 100. Testisuuretta verrataan siis t_{100}-jakaumaan. Sen 95. persentiili on $1,660$.[18] Koska $|0,136| < |1,660|$, niin

[15] Anssi Keinänen ja Auri Pakarinen (2009): Palkkasyrjinnän todistaminen tilastollisesti. Edilex 2009/5 (www.edilex.fi/lakikirjasto/5798.pdf; viitattu 27.4.2011).

[16] Keinänen ja Pakarinen eivät selitä, miten tämä muuttuja on luotu. Selitys yllä on tämän tekstin kirjoittajan päättelemä. Keinänen ja Pakarinen eivät raportoi F-testisuuretta, mutta se on laskettavissa artikkelin tietojen perusteella.

[17] Monilla tilasto-ohjelmistoilla voitaisiin laskea $F_{4,132}$-jakauman tarkka 95. persentiili $(2,440)$.

[18] Useista tilasto-ohjelmistoista saa 95. persentiilin $(1,656)$ t_{132}-jakaumalle.

Standardinormaalijakauman 95. persentiili on $1,645$. Suuren havaintomäärän eli vapausasteiden suuruuden johdosta t- ja standardinormaalijakaumien persentiilit poikkeavat vain vähän toisistaan.

nollahypoteesia ei hylätä 10 %:n riskitasolla. Aineiston mukaan ei ole syytä luopua oletuksesta, että miehet ja naiset saavat samaa palkkaa (kun muut palkkaan vaikuttavat tekijät on huomioitu) eli että palkkasyrjintää ei ole. Ikämuuttujan estimoitu kerroin on $0,001$, ja sen t-arvo on $0,001/0,983 \approx 0,001$. Testisuureen arvon perusteella on selvää, että ikämuuttuja ei voi olla tilastollisesti merkitsevä selittäjä millään järkevällä riskitasolla. Aineiston mukaan ikä ei vaikuta siivojan palkkaan. Samoin voidaan päätellä, että työsuhteen laatu ei näytä vaikuttavan palkkaan ($0,169/0,747 \approx 0,226$). Työsuhteen kesto on tilastollisesti merkitsevä selittäjä palkalle: mallin raportointitarkkuuden yllä mukaan $0,339/0,000 \approx \infty$ (kertoimen estimaatin keskihajonta todellisuudessa on varmasti hieman nollaa suurempi).

Kokeilluista selittäjistä ainoastaan työsuhteen kesto näyttää vaikuttavan siivoojien palkkaan. Kukin työvuosi nostaa palkkaa $0,339$ euroa.

Tutkimuksessa seuraava vaihe voisi olla estimoida malli, jossa ainoa selittäjä on työsuhteen kesto. Tällöin saataisiin luultavasti hieman eri ja hieman tarkempi estimaatti lisätyövuoden palkkaa nostavalle vaikutukselle. Tällaisen mallin vakio olisi palkka siivoojalle, joka on juuri aloittanut siivoojan työuransa.

Mallin (12) vakiolla ei ole järkevää tulkintaa. Kirjaimellisesti tulkiten se olisi palkka 0-vuotiaalle naissiivoojalle, jonka työsuhde on vakinainen mutta jonka työsuhde on vasta alkanut!

3.3.3 Luottamus yhteiskuntaan

Kurssin European Social Survey 2010 (ESS) -osa-aineistossa on viisi järjestysasteikollista $(0, 1, \ldots, 10)$ muuttujaa, jotka kuvaavat suomalaisten luottamusta kansallisiin yhteiskunnallisiin toimijoihin (eduskunta, oikeusjärjestelmä, poliisi, poliitikot ja puolueet). Havainnoista poistettiin 98 vastausta, joissa ei oltu vastattu kaikkiin kysymyksiin. Havaintoja jäi 1780.

Kokeillaan, voidaanko ESS-aineiston ikä- (x_1) ja sukupuolimuuttujilla (x_2) selittää suomalaisten kokemaa luottamusta. Sukupuolimuuttuja saa arvon 0, kun vastaaja on nainen ja muuten arvon 1 (alunperin koodattu ESS-aineistossa hieman toisin).

Koska luottamusmuuttujat ovat diskreettejä, ei jaksojen 3.1.2 ja 3.2.2 oletus regressiomallin jäännöksen normaalijakautuneisuudesta voisi päteä regressiomallissa, jossa luottamusmuuttuja olisi selitettävänä. Niistä laskettiin siksi keskiarvomuuttuja (y). Se voi periaatteessa saada kaikki arvot 0:n ja 10:n välillä 0,2:n välein. Keskiarvomuuttujaa selitettäessä jäännöksen jakauma lienee paremmin approksimoitavissa normaalijakaumalla kuin jotakin alkuperäisistä luottamusmuuttujista selitettäessä. Keskiarvomuuttuja mittaa suomalaisten luottamusta yhteiskuntaan ylipäänsä.

Kukin havaintovektori $[x_{i1} \ \ldots \ x_{i2} \ y_i]$ liittyy yhteen vastaajaan ($i = 1, \ldots, 1780$). PNS-estimointi tuottaa mallin

$$\begin{array}{rcccccccc} y & = & 6,093 & - & 0,003x_1 & - & 0,129x_2 & + & \hat{\varepsilon}. \\ & & (0,114) & & (0,002) & & (0,078) & & \end{array}$$

$$\mathsf{R}^2 = 0,003,\ F_{2,1777} = 2,735,\ n = 1780.$$

Merkinnät ovat kuten edellä.

Jäännös ei ole nyt normaalijakautunut, joten testisuureet eivät ilmeisesti noudata t- ja F-jakaumia. Käytetään niitä karkeina approksimatiivisina vertailujakaumina.

F-testisuure on $F = 2,735$. Verrataan sitä $F_{2,\infty}$-jakauman 95. persentiiliin 2,996 oheisessa taulukossa. Testisuure ei ole sitä suurempi, joten 5 %:n riskitasolla nollahypoteesia (selittäjien kertoimet ovat nollia) ei voi hylätä. R-ohjelmisto raportoi automaattisesti testisuureen p-arvon 0,065. Lähellä hylkäämistä 5 %:n riskitasolla siis ollaan. Selitysosuus on erittäin pieni (0,003), mutta suuren havaintomäärän takia selittäjien pienikin selityskyky voi olla tilastollisesti merkitsevää. Kummankaan selittäjän kertoimen t-arvo ei ole tilastollisesti merkitsevä tavanomaisilla riskitasoilla.

Testauksessa ei ole ylipäänsä syytä ripustautua yhteen tiettyyn riskitasoon. Tämän mallin kohdalla testisuureelle laskettuun p-arvoon tulee suhtautua erityisen varovasti, koska malli ei täytä tavanomaisia oletuksia. Joku saattaisi siksi arvioida, että aineisto varovasti puoltaa nollahypoteesin (kertoimet nollia) hylkäämistä. Tällöin ikä ja sukupuoli selittäisivät suomalaisten kokemaa luottamusta yhteiskuntaan ylipäänsä. Mallin mukaan luottamus heikkenisi iän kasvaessa ja miehet luottaisivat yhteiskuntaan naisia vähemmän. Iän ja sukupuolen vaikutukset olisivat pieniä: Kukin ikävuosi pienentäisi mallin ennustamaa luottamusta $0,003$ yksiköllä. Miehet luottaisivat yhteiskuntaan 0,129 yksikköä vähemmän kuin samanikäiset naiset. Vakiolla ei olisi mielekästä tulkintaa. Kirjaimellisesti se kuvaisi 0-vuotiaan naisen luottamusta yhteiskuntaan.