Yleistetyt lineaariset mallit I

Sara Taskinen ja Arto Luoma
sara.taskinen@jyu.fi, arto.o.luoma@jyu.fi

Syksy 2020

Opintojaksoinfoa

Opintojakso toteutetaan syksyllä 2020 poikkeuksellisesti hybridiopetuksena. Voit suorittaa opintojakson täysin etänä, tai osallistua viikottaisiin Zoom-istuntoihin ja/tai MaD-rakennuksessa järjestettäville kontaktitunneille.

Viikkotehtävät

Opintojakson sisältöä käydään läpi viikottaisten R-tehtävien avulla. Tehtävät ja niihin liittyvät lyhyet ohjevideot avautuvat perjantaina (4.9. alkaen) klo 18:00, jonka jälkeen tehtäviä tehdään itsenäisesti. Tehtäviin on mahdollista kysyä vinkkejä tiistain tai keskiviikon Zoom- tai kontaktitunneilla. Voit ilmoittautua vain yhteen kontaktituntiryhmään! Muistathan, että sairaana ei saa tulla kontaktitunneille.

Kontaktitunnit (8.9. alkaen)

- Ti klo 10:15 - 12 (2 ryhmää)
- Ke klo 14:15 - 16 (2 ryhmää)

Zoom-tunnit (8.9. alkaen)

- Ti klo 16:15 - 18 (1 ryhmä)
- Ke klo 12:15 - 14 (1 ryhmä)

R-harjoitustehtävät (demotehtävät)

Viikottaiset R-harjoitustehtävät (demot) vaikuttavat opintojakson arvolauseeseen. Harjoitustehtävät tehdään itsenäisesti, ja ne arvioidaan itsearviointina malliratkaisujen ja arviointiohjeiden avulla. Opettajalla on mahdollisuus muuttaa itsearvioinnin tulosta.

Harjoitustehtävät avautuvat (9.9. alkaen)

- Ke klo 16:00

Harjoitustehtävät palautetaan

- Pe klo 18:00

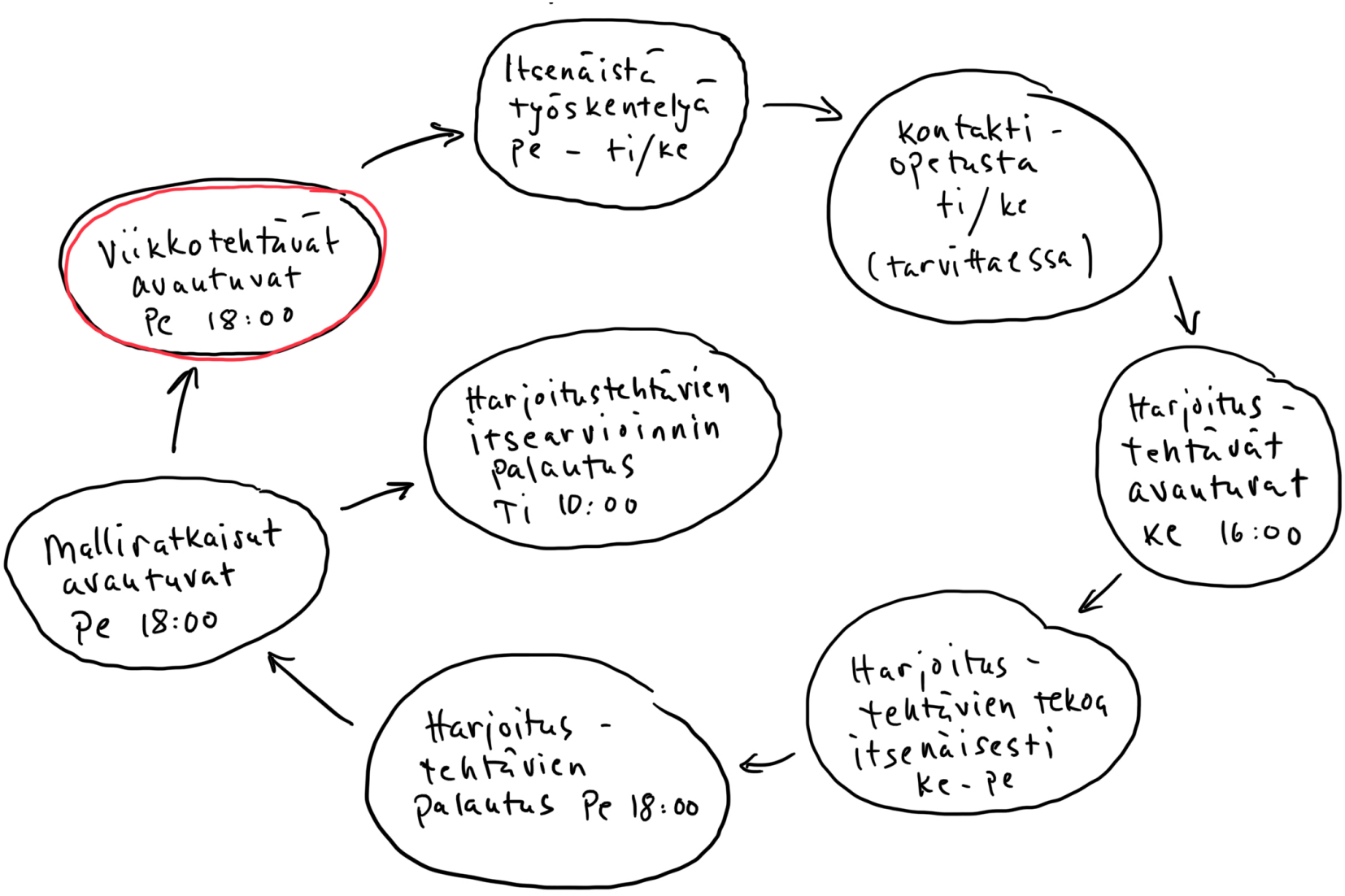
Malliratkaisut ja arviointiohjeet avautuvat

- Pe klo 18:00

Itsevirointi palautetaan

- Ti klo 10:00

Viikkoaikataulu



Opintojakson kuvaus ja esitiedot

- Opintojakso käsittelee yleistettyjen lineaaristen mallien erikoistapauksia, joissa vaste on jatkuva (tavallinen lineaarinen regressio), dikotominen (logistinen regressio) tai lukumäärävaste (Poisson-regressio). Opintojaksolla keskitytään näiden menetelmien ymmärtämiseen. Opintojaksolla opiskellaan R:n (tilastollinen ohjelmointikieli) käyttöä mallintamisessa, laskennassa ja grafiikassa.
- Opintojakso on pääaineopiskelijoiden 2. vuoden aineopintotason opintojakso. Harjoitustehtävien suorittamiseksi vaaditaan R-ohjelmiston käyttötaitoa (TILA410). Tilastotieteen perusopintojen suorittaminen ennen opintojaksoa on toivottavaa. Opintojaksolla käsiteltyjen menetelmien teoriaa käydään yksityiskohtaisesti läpi kevään opintojaksolla Yleistetyt lineaariset mallit 2 (TILA312).

Oppimistavoitteet

Opintojakson menestyksellisesti suorittanut opiskelija osaa:

- tunnistaa kurssilla käsitellyt regressiomallit ja niiden olettamukset,
- antaa empiirisen esimerkin kustakin mallista,
- valita dataan edellä mainituista malleista sopivan ja osaa perustella valintansa vastemuuttujan ominaisuuksilla,
- tehdä tarvittavat numeeriset laskut tilastollisella valmisohjelmalla,
- tehdä empiiriseen aineistoon liittyviä johtopäätöksiä tilastollisen analyysin perusteella,
- arvioida johtopäätöstensä luotettavuutta,
- kirjoittaa perustellun raportin tekemästään tilastollisesta analyysistä.

Esimerkki: Äidin iän ja lapsen syntymäpainon välinen yhteys?

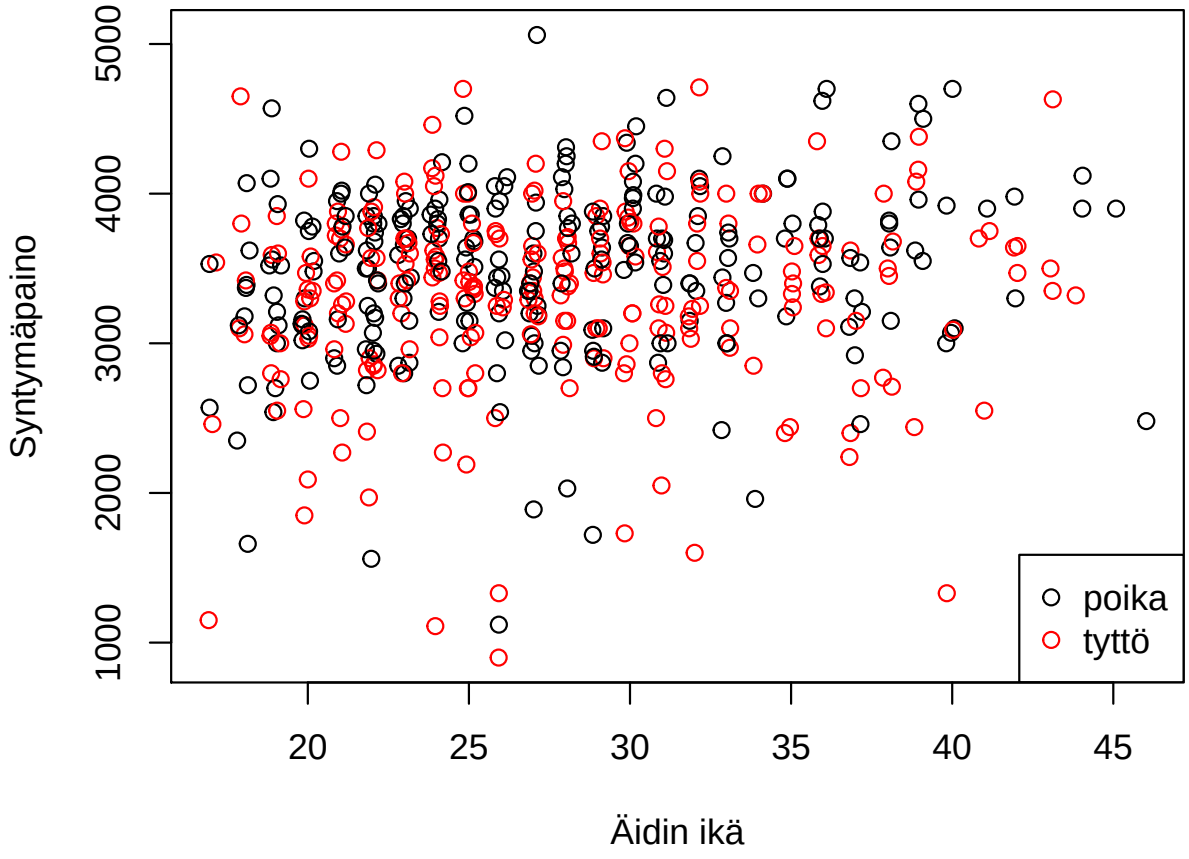


Figure 1: Syntymäpainon ja äidin iän yhteisjakauma.

Mallin valinta, määrittely ja mallioletusten listaaminen

- Tutkimusongelma: Onko yhteys äidin iän ja lapsen syntymäpainon välillä samanlainen tytöillä ja pojilla?
- Sovitetaan aineistoon lineaarinen regressiomalli.

$$paino = \beta_0 + \beta_1\ ika + \beta_2\ sukup + \beta_3\ ika \cdot sukup + \epsilon,$$

missä $\epsilon \sim N(0, \sigma^2)$.

Mallin sovittaminen R ohjelmistolla

```r
dat <- read.table("http://users.jyu.fi/~slahola/files/ylm1_datoja/lapse
                  header = TRUE)
dat$sukup <- factor(dat$sukup)
malli <- lm(paino ~ sukup * ika, data = dat)
summary(malli)
```

```
##
## Call:
## lm(formula = paino ~ sukup * ika, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2420.14  -288.15    75.15   375.53  1563.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3108.145    163.436  19.017   <2e-16 ***
## sukup2        16.111    235.352   0.068   0.9454
## ika           14.392      5.785   2.488   0.0132 *
## sukup2:ika    -6.858      8.319  -0.824   0.4101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Johtopäätösten tekeminen tulosten perusteella

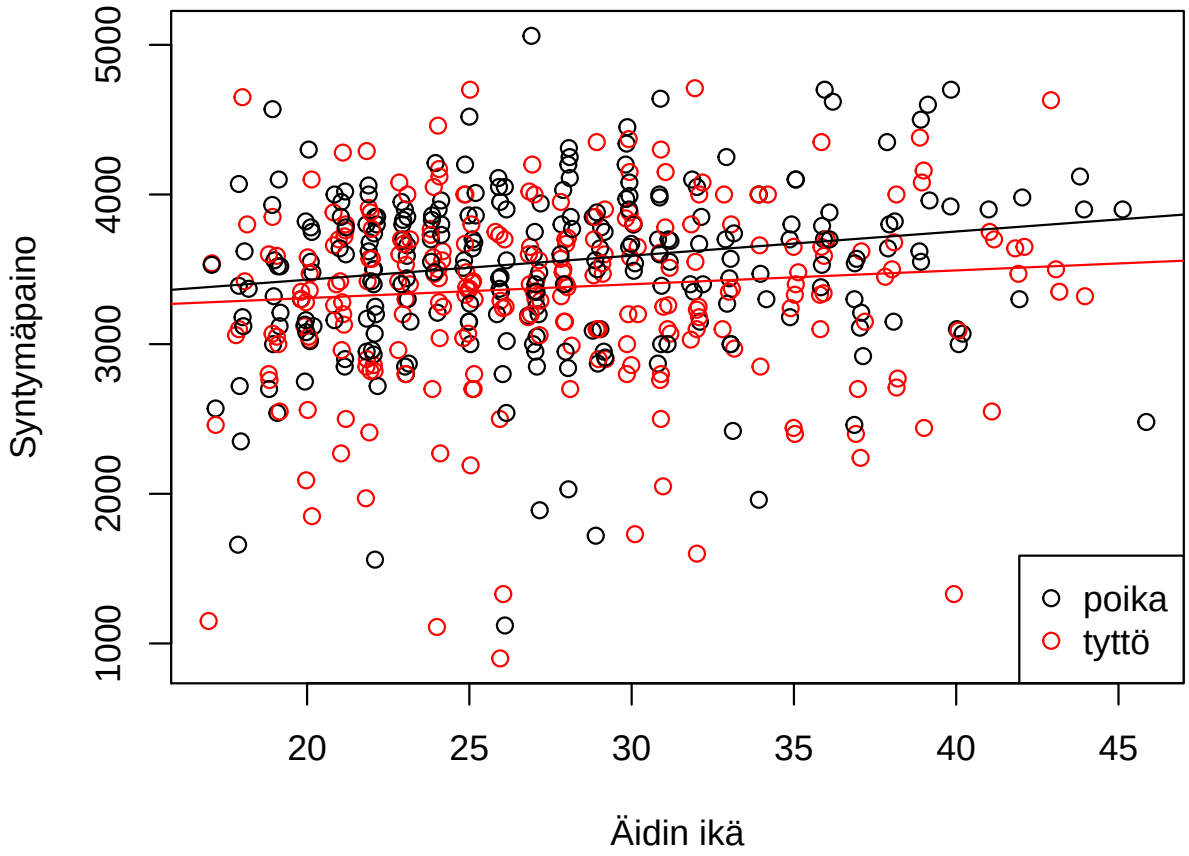


Figure 2: Syntymäpainon ja äidin iän yhteisjakauma sekä estimoidut regressiosuorat.

Johtopäätösten luotettavuuden tutkiminen

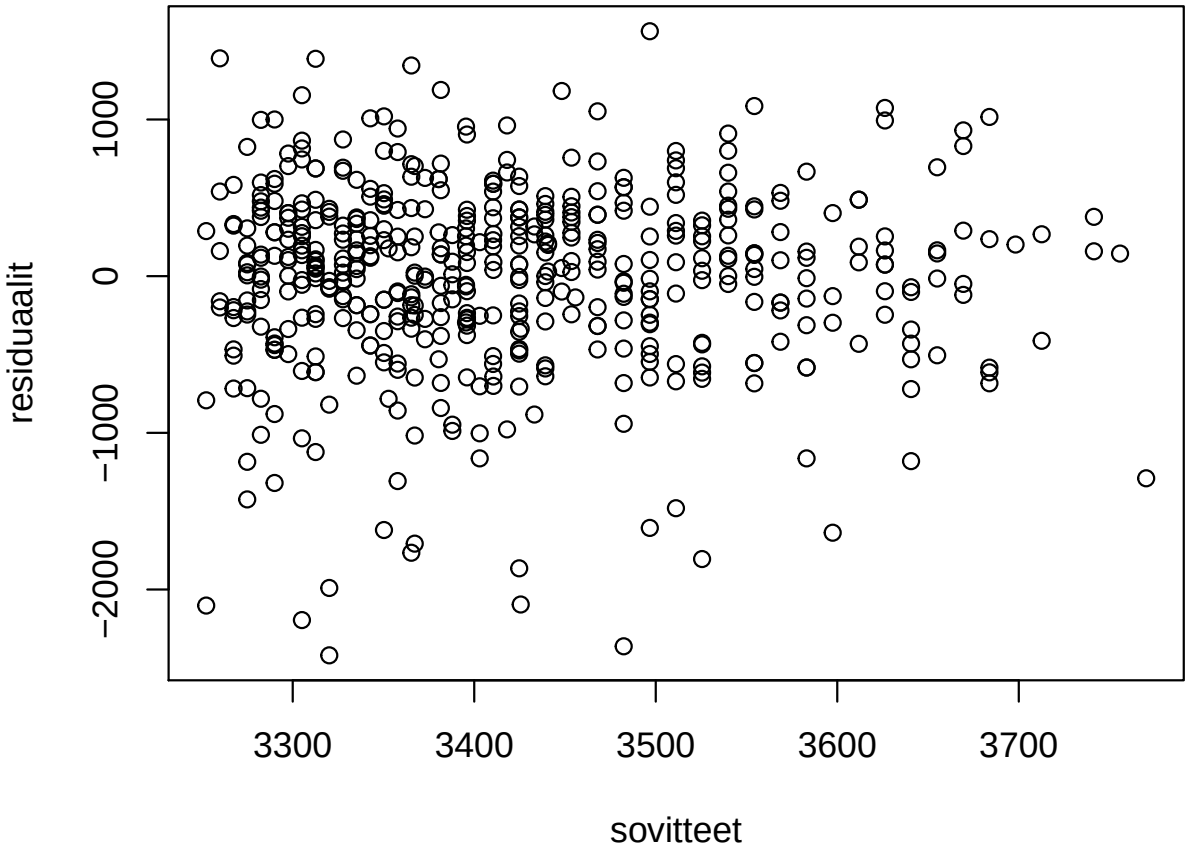
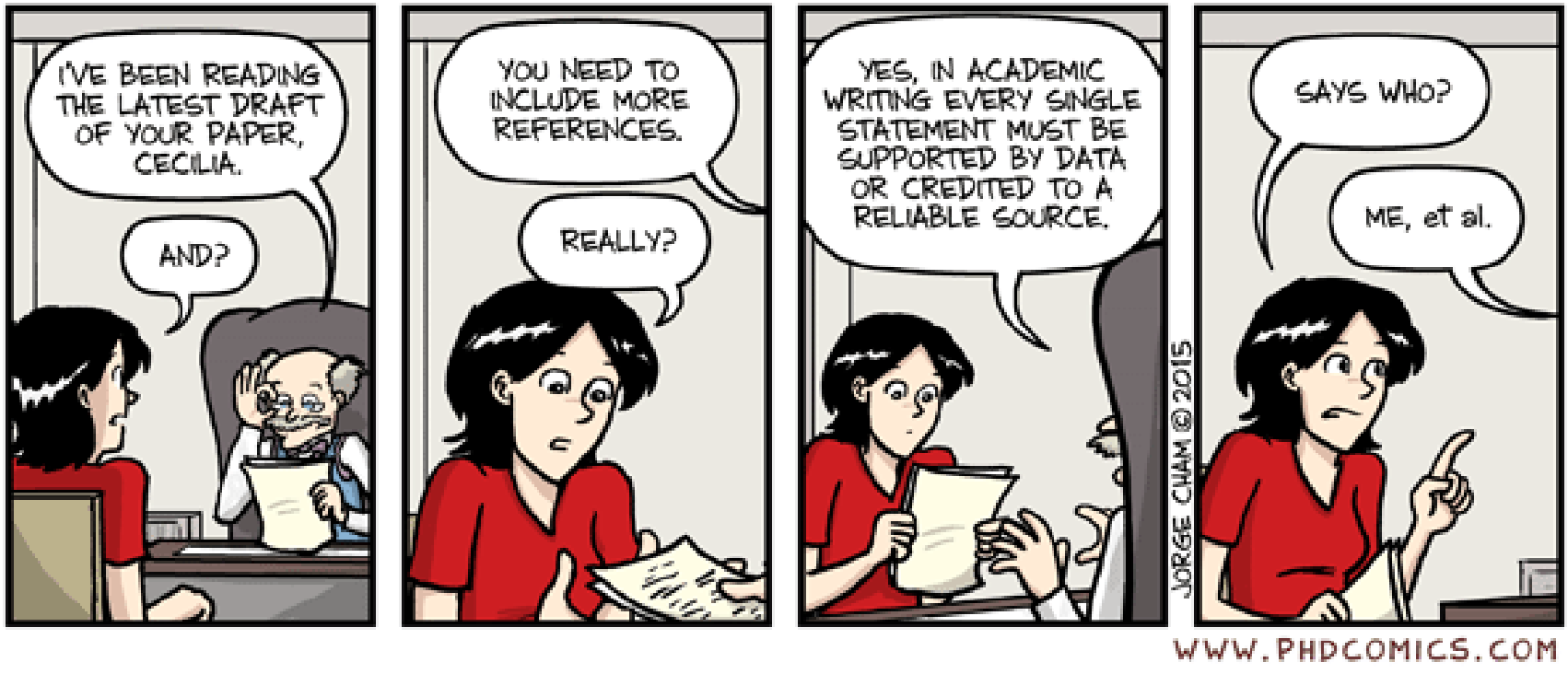


Figure 3: Jäännösten ja sovitteiden yhteisjakauma.

Raportointi



Opintojakson suoritus 5 op:n suorittajille

- Viikottaisten R-tehtävien tekeminen (2 tehtäväsettiä/viikko).
- R harjoitustehtävät (5 kpl), jotka palautetaan Moodleen. Tehtävät itsearvioidaan ja skaalataan opintojakson loppuarvioinnissa siten, että tehtävistä on mahdollista saada maksimissaan 10 pistettä.
- Harjoitustyöt 2 kpl (max 10 pistettä/työ)

Arvosteluasteikko:

- 15 p $\rightarrow$ 1
- 18 p $\rightarrow$ 2
- 21 p $\rightarrow$ 3
- 24 p $\rightarrow$ 4
- 27 p $\rightarrow$ 5

Opintojakson suoritus 2 op:n suorittajille

- Viikottaisten R-tehtävien tekeminen (2 tehtäväsettiä/viikko).
- R harjoitustehtävät (3 kpl), jotka palautetaan Moodleen. Tehtävät itsearvioidaan ja skaalataan opintojakson loppuarvioinnissa siten, että tehtävistä on mahdollista saada maksimissaan 6 pistettä.
- Harjoitustyö (max 10 pistettä)

Arvosteluasteikko:

- 8 p $\rightarrow$ 1
- 9.5 p $\rightarrow$ 2
- 11 p $\rightarrow$ 3
- 12.5 p $\rightarrow$ 4
- 14 p $\rightarrow$ 5

Materiaalia

- Viikkotehtävät perustuvat Jukka Nyblomin (2015) luentomonisteen kappaleisiin 1-4 ja luentokalvoihin. Huomaa, että viikkotehtävät seuraavat hyvin tarkasti luentokalvoja.
- Materiaali on ladattavissa Moodlesta:

  https://moodle.jyu.fi/

- Aineistot ladataan suoraan R:ään verkosta ao. kotisivuilta, joten niitä ei tarvitse tallettaa omalle koneelle.

  http://users.jyu.fi/~slahola/ylm1.htm

- Harjoitustehtävät ja harjoitustyöt palautetaan Moodleen.
- Vinkkejä R-ohjelmiston asennukseen ja käyttöönottoon löydät Moodlen R-oppaasta (lyhyet asennusohjeet löytyvät alta).

Kirjallisuutta

- Gelman and Hill (2007): *Data Analysis using Regression and Multilevel Hierarchical Model.* Cambridge University Press.
- Dobson and Barnett (2008): *An Introduction to Generalized Linear Models.* Chapman and Hall.
- Krzanowski (1998): *An Introduction to Statistical Modelling.* Oxford University Press.

Alustava aikataulu

vko	1. viikkotehtävät	2. viikkotehtävät
1	R alkeet	Lineaarinen regressio, johdanto
2	Useita selittäjiä, interaktio	Estimointi ja inferenssi
3	Diagnostiikka, muunnokset	Ryhmien vertailu
4	Logistinen regressio, johdanto	Useita selittäjiä
5	Interaktio	Inferenssi, diagnostiikka
6	Lukumäärävasteen regressio	Useita selittäjiä, interaktio

R ohjelmiston asentaminen Windows koneelle (ks. Mac/Linux -ohjeet ao. linkistä)

- Mene sivuille: http://cran.r-project.org/index.html
- Klikkaa: “Download R for Windows”
- Klikkaa: “install R for the first time”
- Klikkaa: “Download R X.X.X for Windows”
- Tallenna aukeava tiedosto omalle tietokoneelle.
- Avaa tallentamasi tiedosto (tämä vaihe vaatii riittäviä käyttöoikeuksia).
- Noudata asennusohjelman ohjeita vaihe vaiheelta:
  - Valitse kieli.
  - Voit valita oman asennushakemiston ja muita asetuksia, tai klikata jokaisessa vaiheessa vain “Seuraava/Next”.
  - Voit mm. valita, tuleeko R-kuvake työpöydälle, josta ohjelma aukeaa.
- Avaa R.

RStudion asentaminen

- RStudio on R:n perusversiota käyttäjäystävällisempi.
- Ilmaisen RStudio ohjelmiston (FREE RStudio Desktop) voit ladata sivuilta

https://www.rstudio.com/products/rstudio/download/

- Asennus vaatii R ohjelmiston perusversion asentamisen edellisen sivujen ohjeiden mukaisesti.
- RStudiota voi käyttää tilastollisten menetelmien lisäksi myös moniin muihin tarkoituksiin (nettisovellukset, dokumentit, yms.).