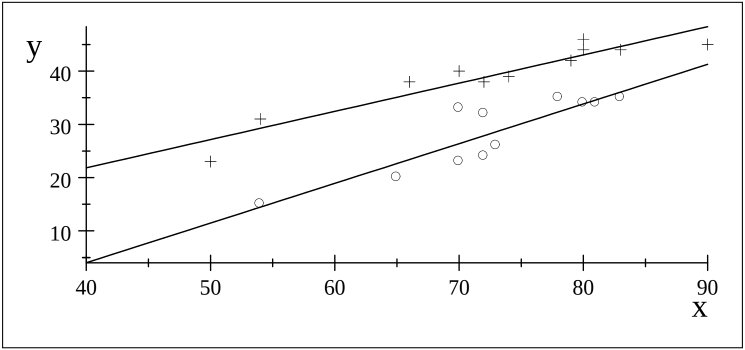
Empiirinen esimerkki

- **Aineisto**: 22 satunnaisesti valittua karitsaa
- **Koeryhmä**: $i = 1, ..., 11$; uusi ruokavalio
- **Kontrolliryhmä**: $i = 11, ..., 22$; vanha ruokavalio
- **Kysymys**: Onko ruokavalioiden välillä eroa ja riippuuko mahdollinen ero ruokinta-ajan pituudesta, joka vaihteli 50 ja 90 päivän välillä?
- **Selitettävä muuttuja** y: Karitsan painon nousu koeajan aikana (mitattuna nauloissa)
- **Selittävät muuttujat**: Koeajan pituus x sekä ryhmää osoittavat indikaattoritmuuttujat d_1 (koeryhmä) ja d_2 (kontrolliryhmä)



Kuva 5.1. Karitsojen painonnousuun (y) ja koeajan pituuteen (x) liittyvä aineisto koeryhmän $(+)$ ja kontrolliryhmän (o) mukaan luokiteltuna sekä kummallekin ryhmälle PNS:llä estimoidut regressiosuorat

$$\text{Koeryhmä:} \qquad y_i = \underset{(4.68)}{0.64} + \underset{(0.064)}{0.53}\, x_i + \hat{\varepsilon}_i, \quad s_1^2 = 5.98$$

$$\text{Kontrolliryhmä:} \qquad y_i = \underset{(9.78)}{-25.80} + \underset{(0.13)}{0.75}\, x_i + \hat{\varepsilon}_i, \quad s_2^2 = 12.23.$$

Empiirinen esimerkki

Koeryhmä: $$y_i = \underset{(4.68)}{0.64} + \underset{(0.064)}{0.53}\, x_i + \hat{\varepsilon}_i, \quad s_1^2 = 5.98$$

Kontrolliryhmä: $$y_i = \underset{(9.78)}{-25.80} + \underset{(0.13)}{0.75}\, x_i + \hat{\varepsilon}_i, \quad s_2^2 = 12.23.$$

- **Kiinnostavia kysymyksiä**: Ovatko regressiosuorat yhdensuuntaisia eli voidaanko estimaattien 0.53 ja 0.75 välinen ero tulkita pelkästään satunnaisvaihtelusta johtuvaksi.
- Jos voidaan, on ruokavalion vaikutuksen mahdollinen ero riippumaton ruokinta-ajan pituudesta, jolloin regressiosuorien vakioiden ero mittaa sitä kaikilla selittävän muuttujan arvoilla.
- Jos regressiosuoria voidaan pitää yhdensuuntaisina, on seuraava kiinnostava kysymys siten voidaanko regressiosuorien vakioiden 0.64 ja -25.80 ero tulkita pelkästään satunnaisvaihtelusta johtuvaksi.

Empiirinen esimerkki

- Kiinnostavien kysymysten tutkiminen perustetaan lineaariseen malliin

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{11} \\ Y_{12} \\ \vdots \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{11} & 0 \\ 0 & 1 & 0 & x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{22} \end{bmatrix},$$

  jossa $\varepsilon_1, ..., \varepsilon_{22} \perp\!\!\!\perp N\left(0, \sigma^2\right)$.
- PNS $\Rightarrow \hat{\alpha}_1 = 0.64$, $\hat{\gamma}_1 = 0.53$, $\hat{\alpha}_2 = -25.80$, $\hat{\gamma}_2 = 0.75$ ja $s^2 = 9.10$ (vrt. aikaisempi).
- Ensimmäinen testattava hypoteesi on $\gamma_1 = \gamma_2$ ja, jos se jää voimaan, testataan hypoteesia $\alpha_1 = \alpha_2$.
- Ennen testaamista on kuitenkin syytä tutkia ovatko mallista tehdyt oletukset realistiset.

Mallin oletusten tarkistaminen

- Mallin oletusten tarkistamisessa residuaalit ovat keskeisiä. Tarkastellaan residuaalivektoria

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\beta}, \qquad \hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y},$$

  jolla on esitys

$$\hat{\boldsymbol{\varepsilon}} = \left(\mathbf{I}_n - \mathbf{P}\right)\mathbf{y}, \quad \mathbf{P} = \left[p_{ij}\right] = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \ \ \text{(projektio)}.$$

- Mallin ollessa oikein spesifioitu pätee vastaavalle sv:lle (HT 3.1)

$$\mathrm{E}\left(\hat{\boldsymbol{\varepsilon}}\right) = \mathbf{0} \quad \text{ja} \quad \mathrm{Cov}\left(\hat{\boldsymbol{\varepsilon}}\right) = \sigma^2\left(\mathbf{I}_n - \mathbf{P}\right) \neq \sigma^2\mathbf{I}_n = \mathrm{Cov}\left(\boldsymbol{\varepsilon}\right).$$

- Residuaalit siis oikeankin mallin kyseessä ollessa korreloituneita ja niiden varianssit vaihtelevat havaintoyksiköstä toiseen. Lisäksi, jos normaalisuusoletus $\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$ pätee, noudattaa residuaalivektori singulaarista normaalijakaumaa $\left(\mathrm{r}\left(\mathrm{Cov}\left(\hat{\boldsymbol{\varepsilon}}\right)\right) = n - p\right)$.

Mallin oletusten tarkistaminen

- Koska $\mathrm{Var}\left(\hat{\varepsilon}_i\right) = \sigma^2(1 - p_{ii})$, on luontevaa tarkastella standardoituja residuaaleja

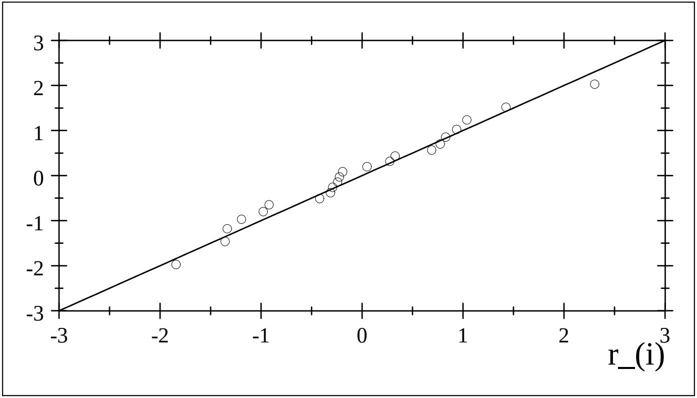
$$r_i = \hat{\varepsilon}_i / s\sqrt{1 - p_{ii}}, \qquad s^2 = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}},$$

  joiden varianssi on likimain 1.
- Järjestetään standardoidut residuaalit $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$.
- Jos $\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$, pätee ($\Phi^{-1}$ on $\mathrm{N}\left(0, 1\right)$-jakauman kf:n käänteisfunktio)

$$\mathrm{E}(R_{(i)}) \approx \Phi^{-1}\left(\frac{i - 0.5}{n}\right) \approx r_{(i)}.$$

- Normaalisuusoletusta voidaan siis tutkia tutkimalla ovatko pisteet $\left(r_{(i)}, \Phi^{-1}\left(\frac{i-0.5}{n}\right)\right)$, $i = 1, ..., n$, likimain 45°:n suoralla.

Mallin oletusten tarkistaminen



Kuva 5.2. Mallin järjestetyt standardoidut residuaalit arvoja $\Phi^{-1}((i-0.5)/n)$ $(i = 1, ..., 22)$ vastaan eli ns. normaalipaperipiirros (normal probability plot tai normal QQ plot).

Tämän perusteella normaalisuusoletus näyttää kohtuulliselta.

Mallin oletusten tarkistaminen

- Muista oletuksista riippumattomuusoletus tuntuu perustellulta, koska kysymyksessä on satunnaisotanta, mutta vaihteleeko Var (ε_i) ryhmien mukaan?
- Koska virheiden $\varepsilon_1, \dots, \varepsilon_{22}$ riippumattomuus voidaan olettaa, on $S_1^2 \perp\!\!\!\perp S_2^2$ ja (Lause 2.1(ii))

$$\frac{S_1^2}{S_2^2} \overset{H_0}{\sim} \frac{1}{9}\chi^2_{1,9} \Big/ \frac{1}{9}\chi^2_{2,9} \sim F_{9,9} \qquad \left(\chi^2_{i,9} \sim \chi^2_9 \text{ ja } \chi^2_{1,9} \perp\!\!\!\perp \chi^2_{2,9}\right).$$

Testisuureen arvoksi saadaan

$$\frac{s_1^2}{s_2^2} = \frac{5.98}{12.23} = 0.49$$

ja P-arvoksi $\mathrm{P}\left(F_{9,9} \leq 0.49\right) = 0.15$. Vakiovarianssioletus tuntuu siten kohtuulliselta.

Tilastollinen analyysi

- Rajoittamaton malli (malli (1)) on

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{11} \\ Y_{12} \\ \vdots \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{11} & 0 \\ 0 & 1 & 0 & x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{22} \end{bmatrix}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right),$$

josta estimaatit $\hat{\alpha}_1 = 0.64$, $\hat{\gamma}_1 = 0.53$, $\hat{\alpha}_2 = -25.80$, ja $\hat{\gamma}_2 = 0.75$.

- Tarkastellaan hypoteesia $H_0 : \gamma_1 = \gamma_2 = \gamma$.

Tilastollinen analyysi

- Rajoitettu malli on

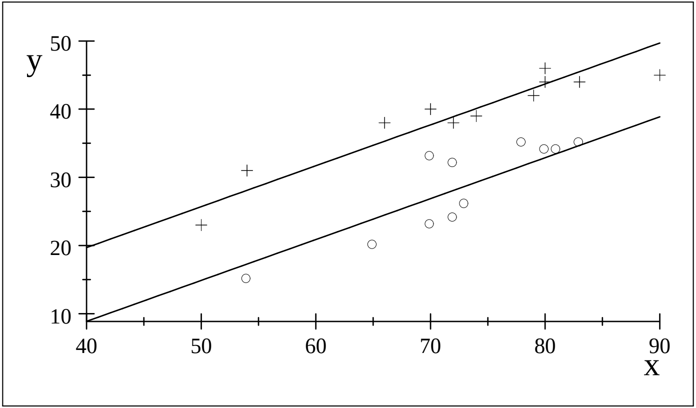
$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{11} \\ Y_{12} \\ \vdots \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{11} \\ 0 & 1 & x_{12} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{22} \end{bmatrix}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right).$$

- Virhevarianssin estimaatiksi tulee $s_H^2 = 9.76$ ja $n - p_H = 22 - 3 = 19$.
- Vapaasta mallista saatiin $s^2 = 9.10$ ja $n - p = 22 - 4 = 18$.
- Testisuure saa arvon

$$F = \frac{(19 \times 9.76 - 18 \times 9.10)/1}{9.10} = 2.37$$

ja vastaavaksi P-arvoksi tulee

$$\mathrm{P} = \mathrm{P}\left(F_{1,18} \geq 2.37\right) = 0.14 \;\Rightarrow\; H_0 : \gamma_1 = \gamma_2 \text{ jää voimaan.}$$



Kuva 5.3. Kuvan 5.1 aineistoon piirretyt mallin (2) estimointituloksiin perustuvat samansuuntaiset regressiosuorat.

Estimoitu malli

$$y_i = \underset{(4.98)}{-4.30}\, d_{i1} - \underset{(4.98)}{15.12} d_{i2} + \underset{(0.067)}{0.60}\, x_i + \hat{\varepsilon}_i, \quad s^2 = 9.76 \qquad (2)$$

Tilastollinen analyysi

- Tarkastellaan nyt hypoteesia $H_0 : \alpha_1 = \alpha_2 = \alpha$ mallissa (2)

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{11} \\ Y_{12} \\ \vdots \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{11} \\ 0 & 1 & x_{12} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{22} \end{bmatrix}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right).$$

- Rajoitettu malli on

$$Y_i = \alpha + \gamma x_i + \varepsilon_i, \quad i = 1, ..., 22.$$

PNS-estimointi $\Rightarrow$

$$y_i = \underset{(10.16)}{-9.71} + \underset{(0.14)}{0.60}\, x_i + \hat{\varepsilon}_i, \qquad i = 1, ..., 22, \qquad s^2 = 41.46 \qquad (3)$$

Tilastollinen analyysi

- Vapaassa mallissa (2)

$$s^2 = 9.76, \quad n - p = 22 - 3 = 19$$

- Rajoitetussa mallissa (3)

$$s^2 = 41.46, \quad n - p_H = 22 - 2 = 20$$

- Testisuure saa arvon

$$F = \frac{(20 \times 41.46 - 19 \times 9.76)/1}{9.76} = 65.97$$

ja vastaavaksi P-arvoksi tulee

$$\mathrm{P} = \mathrm{P}(F_{1,19} \geq 65.97) \approx 0 \; \Rightarrow H_0 : \alpha_1 = \alpha_2 \text{ on syytä hylätä.}$$

Tilastollinen analyysi

- Malli (2)

$$Y_i = \alpha_1 d_{i1} + \alpha_2 d_{i2} + \gamma x_i + \varepsilon_i, \quad i = 1, ..., 22.$$

- Ruokavalioiden eroa mittaa parametri

$$\alpha_1 - \alpha_2 = [1 \quad -1 \quad 0] \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma \end{bmatrix} = a'\beta.$$

- $\hat{\alpha}_1 - \hat{\alpha}_2 = -4.30 + 15.12 = 10.82, \quad s\sqrt{\mathbf{a}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{a}} = 1.77,$
  $t_{19}\left(0.05/2\right) = 2.09$

- 95%:n luottamusväli erotukselle $\alpha_1 - \alpha_2$ on

$$10.82 - 2.09 \cdot 1.77 < \alpha_1 - \alpha_2 < 10.82 + 2.09 \cdot 1.77$$

eli

$$7.12 < \alpha_1 - \alpha_2 < 14.52$$

Tilastollinen analyysi

Johtopäätös

- Uusi ruokavalio on (nopean painon nousun kannalta) systemaattisesti vanhaa parempi, sillä sen paremmuus ei riipu tarkasteltujen ruokinta-aikojen pituudesta.
- Kun ruokinta-ajan pituus on n. 50 - 90 päivää, voidaan uudella ruokavaliolla odottaa päästävän 95%:n varmuudella n. 7 - 14.5 naulaa suurempaan karitsan painon nousuun kuin vanhalla ruokavaliolla.

Yksisuuntainen varianssianalyysi

Yksisuuntaisessa varianssianalyysissa tarkastellaan p:tä ryhmää ja niistä satunnaisotannalla poimittuja havaintoja. Mielenkiinnon kohteena on ryhmien odotusarvojen mahdolliset erot. Asetelma on kaaviona

	Havainnot	Keskiarvot
Ryhmä 1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	$\bar{y}_1$
Ryhmä 2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$
Ryhmä p	$y_{p1}, y_{p2}, \ldots, y_{pn_p}$	$\bar{y}_p$

jossa $\bar{y}_j = \left(y_{j1} + \cdots + y_{jn_j} \right) / n_j \; (j = 1, \ldots, p)$.

Yksisuuntainen varianssianalyysi

- Tilastollista mallia varten oletetaan, että havaintoja vastaavat satunnaismuuttujat $Y_{11}, ..., Y_{1n_1}, Y_{21}, ..., Y_{2n_2}, ..., Y_{p1}, ..., Y_{pn_p}$ ovat riippumattomia ja $Y_{ji} \sim \mathrm{N}(\mu_j, \sigma^2)$
- tai yhtäpitävästi, että

$$Y_{ji} = \mu_j + \varepsilon_{ji}, \quad i = 1, ..., n_j, \; j = 1, ..., p,$$

jossa sm:t $\varepsilon_{ji} \sim \mathrm{N}(0, \sigma^2)$ ovat riippumattomia ja $\mu_j \in \mathbb{R}$.

- Käyttäen matriisimerkintöjä saadaan yhtälö

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_p \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_p} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_p \end{bmatrix},$$

josta nähdään, että kysymyksessä on lineaarisen mallin erikoistapaus.

Yksisuuntainen varianssianalyysi

- $Y_{ji} = \mu_j + \varepsilon_{ji}, \quad i = 1, ..., n_j, \; j = 1, ..., p,$

  jossa sm:t $\varepsilon_{ji} \sim \mathrm{N}(0, \sigma^2)$ ovat riippumattomia ja $\mu_j \in \mathbb{R}$.
- Mielenkiinnon kohteena on nollahypoteesi

  $$H : \mu_1 = \cdots = \mu_p \quad \Leftrightarrow \quad \mu_1 - \mu_p = \mu_2 - \mu_p = \cdots = \mu_{p-1} - \mu_p = 0,$$

  tai matriisein

  $$\begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & -1 \\ 0 & 1 & 0 & \cdots & \cdots & -1 \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}.$$
- Nollahypoteesi on siis lineaarista muotoa $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$, jossa $\mathbf{A}$ on $(p-1) \times p$ matriisi ja $\mathrm{r}(\mathbf{A}) = p - 1$.

Yksisuuntainen varianssianalyysi

- Kysymyksessä on lineaarinen malli

$$Y_{ji} = \mu_j + \varepsilon_{ji}, \quad i = 1, ..., n_j, \ j = 1, ..., p, \qquad (*)$$

jossa sm:t $\varepsilon_{ji} \sim \mathrm{N}\left(0, \sigma^2\right)$ ovat riippumattomia ja $\mu_j \in \mathbb{R}$. Testattava (lineaarinen) hypoteesi

$$H : \mu_1 = \cdots = \mu_p := \mu_0.$$

- Testisuureen yleinen lauseke on $(\mathrm{r}\left(\mathbf{A}\right) = p - 1)$

$$F = \frac{\left(\mathrm{S}(\hat{\boldsymbol{\mu}}_H) - \mathrm{S}(\hat{\boldsymbol{\mu}})\right) / \left(p-1\right)}{\mathrm{S}(\hat{\boldsymbol{\mu}}) / \left(n-p\right)} \overset{H}{\sim} F_{p-1,n-p},$$

jossa parametrin $\boldsymbol{\mu} = \left(\mu_1, ..., \mu_p\right)$ vapaa PNS-estimaatti $\hat{\boldsymbol{\mu}}$ saadaan mallista $(*)$ ja sidottu PNS-estimaatti $\hat{\boldsymbol{\mu}}_H$ mallista

$$Y_{ji} = \mu_0 + \varepsilon_{ji}, \quad i = 1, ..., n_j, \ j = 1, ..., p.$$

Yksisuuntainen varianssianalyysi

- Sidottu PNS-estimaatti $\hat{\boldsymbol{\mu}}_H$ minimoi jäännösnelineliösumman

$$\mathrm{S}(\boldsymbol{\mu}_0) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \mu_0)^2 \,.$$

Siten, $\hat{\boldsymbol{\mu}}_H = (\bar{y}, ..., \bar{y})$, jossa

$$\hat{\mu}_0 = \bar{y} = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} y_{ji} = \frac{1}{n} \sum_{j=1}^{p} n_j \bar{y}_j \quad (n = n_1 + \cdots + n_p) \,,$$

ja

$$\mathrm{S}\left(\hat{\boldsymbol{\mu}}_H\right) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \bar{y})^2 \,.$$

- Vapaa PNS-estimaatti $\hat{\boldsymbol{\mu}}$ minimoi jäännösneliöneliösumman

$$\mathrm{S}\left(\boldsymbol{\mu}\right) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} \left(y_{ji} - \mu_j\right)^2 .$$

- Tämä johtaa estimaatteihin $\hat{\mu}_j = \bar{y}_j = \left(y_{j1} + \cdots + y_{jn_j}\right) / n_j$, joten $\hat{\boldsymbol{\mu}} = \left(\bar{y}_1, ..., \bar{y}_p\right)$ ja

$$\mathrm{S}(\hat{\boldsymbol{\mu}}) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} \left(y_{ji} - \bar{y}_j\right)^2$$

Yksisuuntainen varianssianalyysi

- Testisuureeksi saadaan

$$F = \frac{(S(\hat{\mu}_H) - S(\hat{\mu}))/(p-1)}{S(\hat{\mu})/(n-p)}$$

$$= \frac{\left(\sum_{j=1}^{p}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y})^2 - \sum_{j=1}^{p}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2\right)/(p-1)}{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2/(n-p)}$$

$$\overset{H}{\sim} F_{p-1,n-p},$$

- Tämä riittää, mutta testisuure on perinteisesti esitetty hieman toisessa muodossa.

Yksisuuntainen varianssianalyysi

- Laskemalla nähdään, että F-testisuureen osoittaja voidaan kirjoittaa

$$\mathrm{S}\left(\hat{\mu}_H\right) - \mathrm{S}\left(\hat{\mu}\right) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} \left(\bar{y}_j - \bar{y}\right)^2 = \sum_{j=1}^{p} n_j \left(\bar{y}_j - \bar{y}\right)^2,$$

  joten F-testisuureeksi saadaan

$$F = \frac{\sum_{j=1}^{p} n_j \left(\bar{Y}_j - \bar{Y}\right)^2 / (p-1)}{\sum_{j=1}^{p} \sum_{i=1}^{n_j} \left(Y_{ji} - \bar{Y}_j\right)^2 / (n-p)} \overset{H}{\sim} F_{p-1,n-p}.$$

- Usein sanotaan, että testisuure on ryhmien välisen varianssin ja ryhmien sisäisen varianssin suhde, mistä nimitys varianssianalyysi tulee.
- Huomaa, että testisuureen nimittäjässä on $\mathrm{S}\left(\hat{\mu}\right) / (n-p) = S^2$.

Yksisuuntainen varianssianalyysi

- Jos nollahypoteesi hylätään, on kiinnostavaa tutkia hypoteeseja $\mu_{j_1} - \mu_{j_2} = 0$ tai yleisemmin ns. *kontrasteja* $\mathbf{a}'\boldsymbol{\mu} = 0$, jossa $\mathbf{a}' = \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix}$ toteuttaa $a_1 + \cdots + a_p = 0$.
- Koska $\mathbf{X}'\mathbf{X} = \mathrm{diag}\begin{bmatrix} n_1 & \cdots & n_p \end{bmatrix}$, pätee (Lauseen 2.1(i))

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \bar{Y}_1 & \cdots & \bar{Y}_p \end{bmatrix}' \sim \mathrm{N}\left(\boldsymbol{\mu}, \sigma^2 \mathrm{diag}\begin{bmatrix} \frac{1}{n_1} & \cdots & \frac{1}{n_p} \end{bmatrix}\right).$$

- Nollahypoteesin $\mathbf{a}'\boldsymbol{\mu} = 0$ voimassa ollessa pätee siten (Liite A.2.4)

$$\mathbf{a}'\hat{\boldsymbol{\mu}} \sim \mathrm{N}\left(0, \sigma^2 \sum_{j=1}^{p} a_j^2 / n_j\right).$$

- Tämä johtaa F-testisuureeseen

$$F = \frac{\left(\sum_{j=1}^{p} a_j \bar{Y}_j\right)^2}{S^2 \sum_{j=1}^{p} a_j^2 / n_j} \overset{H}{\sim} F_{1,n-p}, \quad S^2 = \frac{1}{n-p} \sum_{j=1}^{p} \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y}_j)^2 .$$

- F-testisuureen

$$F = \frac{\left(\sum_{j=1}^{p} a_j \bar{Y}_j\right)^2}{S^2 \sum_{j=1}^{p} a_j^2 / n_j} \overset{H}{\sim} F_{1,n-p}, \quad S^2 = \frac{1}{n-p} \sum_{j=1}^{p} \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y}_j)^2$$

vaihtoehtona voidaan käyttää testisuuretta (ks. jakso 4.1)

$$T = \frac{\sum_{j=1}^{p} a_j \bar{Y}_j}{S \sqrt{\sum_{j=1}^{p} a_j^2 / n_j}} \overset{H}{\sim} t_{n-p}$$

- Kontrastin $\mathbf{a}'\boldsymbol{\mu}$ luottamusväliksi luottamustasolla $1 - \alpha$ saadaan

$$\sum_{j=1}^{p} a_j \bar{y}_j \pm t_{n-p}(\alpha/2)\, s \sqrt{\sum_{j=1}^{p} a_j^2 / n_j}.$$

- Useita luottamusvälejä muodostettaessa on kuitenkin syytä muistaa jaksossa 4.1 mainitut hankaluudet.