Peatükk 4

F-test

4.1 Mudelite võrdlemine F-testi abil

Olgu meil vaatluse all kaks mudelit — lihtsam mudel $\mathcal{M}_0$ mudeli maatriksiga $\mathbf{X}_0$ ja keerukam mudel $\mathcal{M}_1$ mudeli maatriksiga $\mathbf{X}_1$. Eeldame, et lihtsam mudel on erijuht keerukamast — keerukamast mudelist on võimalik lihtsamat mudelit saada fikseerides osade parameetrite väärtused (näiteks võrdsustades mõne parameetri nulliga). Seega on mudel $\mathcal{M}_0 : y = c_0 + \varepsilon$ erijuht mudelist $\mathcal{M}_1 : y = c_0 + c_1 x + \varepsilon$, sest keerukamast mudelist saaksime lihtsama, kui võtaksime $c_1 = 0$. Matemaatiliselt korrektsemalt kirja pandult: üks mudel on teise erijuht, kui $\mathcal{C}(\mathbf{X}_0) \subset \mathcal{C}(\mathbf{X}_1)$. Meie sooviks on testida, kas lihtsama mudeli kasutamine on õigustatud (nullhüpotees: lihtsam mudel on õige) või sunnivad andmed meid kasutama keerukamat mudelit:

$$\begin{aligned} H_0 : \mathbf{y} &\sim N(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2\mathbf{I}) \\ H_1 : \mathbf{y} &\sim N(\mathbf{X}_1\boldsymbol{\beta}_1, \sigma^2\mathbf{I}) \end{aligned}$$

Paneme tähele: kui lihtsam mudel $\mathcal{M}_0$ on õige, siis on õige ka keerukam mudel $\mathcal{M}_1$. Selgitus: Kui $\mathcal{C}(\mathbf{X}_0) \subset \mathcal{C}(\mathbf{X}_1)$ siis on võimalik leida maatriks $\mathbf{M}$ selliselt, et $\mathbf{X}_0 = \mathbf{X}_1\mathbf{M}$. Kui lihtsam mudel on korrektne, siis valiku $\boldsymbol{\beta}_1 = \mathbf{M}\boldsymbol{\beta}_0$ korral on ka keerukam mudel õige: $\mathbf{X}_1\boldsymbol{\beta}_1 = \mathbf{X}_1\mathbf{M}\boldsymbol{\beta}_0 = \mathbf{X}_0\boldsymbol{\beta}_0$.

Antud peatükis eeldame, et keerukam mudel on igal juhul õige. Meie soovime vaid teada, kas ka lihtsam mudel on õige.

Kuna on üks mudel erijuht teisest? Vaatame paari näidet. Kahes esimeses näites on lihtsam mudel erijuht keerukamast. Viimases näites on tegemist aga mudelitega, milles üks pole vaadeldav teise mudeli lihtsustusena (seega ei saa neid mudeleid võrrelda siin peatükis toodud lähenemise abil).

Näide 4.1 *Kasutada prognoosimisel maakonda või valda?*

Oletame, et oleme mõõtnud inimeste keskmist palka kolmes maakonnas — Harjumaal, Tartumaal ja Hiiumaal. Igas maakonnas oleme reaalselt küsitlenud inimesi kahes omavalitsuses (Harjumaal: Viimsi ja Saku vallas; Tartumaal Tartu linnas ja Konguta vallas; Hiiumaal Kärdla linnas ja Emmaste vallas). Kogutav andmestik võiks välja näha selline:

Palk	*Maakond*	*Vald*
Y_1	*Harjumaa*	*Viimsi*
Y_2	*Harjumaa*	*Viimsi*
Y_3	*Harjumaa*	*Saku*
Y_4	*Harjumaa*	*Saku*
Y_5	*Tartumaa*	*Tartu*
Y_6	*Tartumaa*	*Tartu*
Y_7	*Tartumaa*	*Konguta*
Y_8	*Tartumaa*	*Konguta*
Y_9	*Hiiumaa*	*Kärdla*
Y_{10}	*Hiiumaa*	*Kärdla*
Y_{11}	*Hiiumaa*	*Emmaste*
Y_{12}	*Hiiumaa*	*Emmaste*

Kas inimese palga prognoosimiseks piisab faktortunnuse „maakond“ kasutamisest või saame parema mudeli tunnust „vald“ kasutades? Prognoosimiseks maakonda kasutava mudeli mudelimaatriks $\mathbf{X}_0$ *ja valda kasutava mudeli (keerukama mudeli) mudelimaatriks* $\mathbf{X}_1$ *näevad välja nii:*

$$X_0 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \qquad X_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Kuna $\mathcal{C}(\mathbf{X}_0) \subset \mathcal{C}(\mathbf{X}_1)$ *siis on üks vaadeldav mudel erijuht teisest.*

Näide 4.2 *Kaks mõõtmist — kas kasutada mõlemat või mõõtmiste keskmist?*

Soovime prognoosida laste kaalu. Meil on i. lapse kaalu prognoosimiseks kasutada kaks tunnust:

- *lastearsti poolt mõõdetud lapse pikkus (x_{1i})*
- *lapsevanema poolt väidetav lapse pikkus (x_{2i})*

Valime kahe mudeli vahel. Keerukam mudel kasutab prognoosimiseks nii lapsevanema kui ka lastearsti mõõdetud pikkust:

$$y_i = c_0 + c_1 x_{1i} + c_2 x_{2i} + \varepsilon_i.$$

Lihtsam mudel kasutab aga kaalu prognoosimiseks kahe pikkusemõõtmise keskmist:

$$y_i = c_0^* + c_1^*(x_{1i} + x_{2i})/2 + \varepsilon_i.$$

Toodud mudelite puhul on keerukam mudel lihtsama erijuht, sest valides $c_0 = c_0^*$, $c_1 = c_1^*/2$ ja $c_2 = c_1^*/2$ saame keerukamast mudelist lihtsama mudeli.

Näide 4.3 *Kaks mõõtmist — kas kasutada lastearsti või lapsevanema poolt mõõdetud lapse pikkust? Vaatame eelmises näites kirjeldatud olukorda — prognoosime lapse kaalu pikkuse järgi. Kas peaksime kasutama lastearsti või lapsevanema poolt mõõdetud lapse pikkust? Ehk kas peaksime eelistama mudelit*

$$y_i = c_0 + c_1 x_{1i} + \varepsilon_i$$

või mudelit

$$y_i = c_0 + c_2 x_{2i} + \varepsilon_i?$$

Toodud mudelist ei saa kumbagi vaadelda kui teise mudeli erijuhtu (välja arvatud juhul, kui lapsevanema ja lastearsti mõõtmistulemused langevad täpselt kokku). Seega ei saa nende kahe mudeli seast sobivaimat leida käesoleva peatükis käsitlemist leidavate meetodite abil.

Kuidas otsustada, kas kasutada lihtsamat või keerukamat mudelit? Üldine printsiip (Ockhami habemenoa printsiip) on järgmine — kui pole (piisavalt) tõendusmateriali, mis räägiks keerukama mudeli kasuks, siis eelistame lihtsamat mudelit. Kuidas siis mudelite korral otsustada, kas andmed „sunnivad“ meid keerukamat mudelit kasutama või mitte?

Vaatame mõlema mudeli prognoose vaatlusvektorile, $\hat{\mathbf{y}}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}}_0$ ja $\hat{\mathbf{y}}_1 = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1$. Kui lihtsam mudel on ka õige, siis peaks prognooside vahe $\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0$

tulema väike (nullilähedane), ehk prognooside erinevuste ruutude summa $(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0)^T(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0)$ peaks tulema pisike. Seda ruutude summat saab kirja panna ka veidi teisel moel:

$$\begin{aligned}(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0)^T(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0) &= ((\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y})^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} \\ &= \mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} \\ &= \mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1}\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_1}\mathbf{P}_{\mathbf{X}_0} - \mathbf{P}_{\mathbf{X}_0}\mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{X}_0}\mathbf{P}_{\mathbf{X}_0})\mathbf{y}.\end{aligned}$$

Kuna $\mathcal{C}(\mathbf{X}_0) \subset \mathcal{C}(\mathbf{X}_1)$ siis leidub maatriks $\mathbf{M}$ nii, et $\mathbf{X}_0 = \mathbf{X}_1\mathbf{M}$. Seega

$$\begin{aligned}\mathbf{P}_{\mathbf{X}_1}\mathbf{P}_{\mathbf{X}_0} &= \mathbf{P}_{\mathbf{X}_1}\mathbf{X}_1\mathbf{M}(\mathbf{X}_0{}^T\mathbf{X}_0)^-\mathbf{X}_0{}^T \\ &= \mathbf{X}_1\mathbf{M}(\mathbf{X}_0{}^T\mathbf{X}_0)^-\mathbf{X}_0{}^T \\ &= \mathbf{X}_0(\mathbf{X}_0{}^T\mathbf{X}_0)^-\mathbf{X}_0{}^T \\ &= \mathbf{P}_{\mathbf{X}_0}\end{aligned}$$

ja järelikult

$$(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0)^T(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0) = \mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}.$$

Kui suurt prognooside erinevust võiksime näha, kui mõlemad mudelid on õiged? Sellele küsimusele aitab vastata meile juba tuttav valem (vaata lemmat 3.1): $\mathrm{E}\mathbf{y}^T\mathbf{A}\mathbf{y} = \mathrm{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$. Seda valemit ja tähistusi $p_0 := \mathrm{rank}(\mathbf{X}_0)$ ja $p_1 := \mathrm{rank}(\mathbf{X}_1)$ kasutades võime kirjutada:

$$\begin{aligned}\mathrm{E}\left[\mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}\right] &= \mathrm{tr}\left[(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\sigma^2\right] + \boldsymbol{\beta}_0^T\mathbf{X}_0^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_0\boldsymbol{\beta} \\ &= \sigma^2(p_1 - p_0) + 0,\end{aligned}$$

Sest $\mathbf{X}_0{}^T\mathbf{P}_{\mathbf{X}_1} = \mathbf{X}_0{}^T$ (meenuta: $\mathbf{X}_0^T = \mathbf{M}^T\mathbf{X}_1^T$) ja $\mathbf{X}_0{}^T\mathbf{P}_{\mathbf{X}_0} = \mathbf{X}_0{}^T$. Teisisõnu, nullhüpoteesi kehtides:

$$\mathrm{E}\left[\frac{\mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}}{p_1 - p_0}\right] = \sigma^2.$$

Meenutame, et oli teinegi nihketa hinnang jääkide dispersioonile, MSE:

$$\mathrm{E}\left[\frac{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y}}{n - p_1}\right] = \sigma^2.$$

Viimane hinnang on õige ka siis, kui lihtsam mudel ei kehti (keerukam mudel on aga siiski õige). Seega võiks nende kahe hinnangu suhe nullhüpoteesi kehtides (kui mõlemad mudelid on õiged) olla ligikaudu 1:

$$\frac{\mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}}{p_1 - p_0} \Big/ \frac{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y}}{n - p_1} \overset{H_0}{\approx} 1.$$

Aga kui kaugele ühest võib see suhe puhtalt juhuse tõttu kalduda?

Meenutame üht-teist jaotuste kohta.

Kui $\mathbf{y} \sim N(\boldsymbol{\mu}; \mathbf{I})$ ja $\mathbf{A}$ on sümmeetriline ja idempotentne maatriks ($\mathbf{AA} = \mathbf{A}$), siis $\mathbf{y}^T\mathbf{Ay} \sim {\chi^2}_{df=\text{rank}(\mathbf{A});\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}}$. Seega, nullhüpoteesi kehtides ($\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_0\boldsymbol{\beta}_0$):

$$\begin{aligned}(\mathbf{y}^T/\sigma)(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})(\mathbf{y}/\sigma) &= \mathbf{y}^T(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})\mathbf{y}/\sigma^2 \\ &\sim {\chi^2}_{p_1-p_0;\boldsymbol{\beta}_0^T\mathbf{X}_0{}^T(\mathbf{P}_{\mathbf{X}1}-\mathbf{P}_{\mathbf{X}0})\mathbf{X}_0\boldsymbol{\beta}_0/\sigma^2} \\ &\sim {\chi^2}_{p_1-p_0;0}\end{aligned}$$

ja

$$\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}1})\mathbf{y}/\sigma^2 \sim {\chi^2}_{n-p_1}.$$

Nüüd veel meenutust: kui $X \sim \chi^2_k$ ja $Y \sim \chi^2_l$ on sõltumatud juhuslikud suurused, siis

$$\frac{X/k}{Y/l} \sim F_{k,l}.$$

Järelikult, kui lugeja ja nimetaja oleksid sõltumatud (hetke pärast kontrollime), oleks:

$$\frac{\mathbf{y}^T(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})\mathbf{y}/\left[\sigma^2(p_1 - p_0)\right]}{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}1})\mathbf{y}/\left[\sigma^2(n - p_1)\right]} \overset{H_0}{\sim} F - p_1 - p_0; n - p_1.$$

Kas lugeja ja nimetaja on sõltumatud? Selle näitamiseks piisab, kui saaksime näidata $(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})\mathbf{y}$ ja $(\mathbf{I} - \mathbf{P}_{\mathbf{X}1})\mathbf{y}$ sõltumatust (sest $f(X) \perp g(Y)$ kui $X \perp Y$). Kuna $((\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0}) : (\mathbf{I} - \mathbf{P}_{\mathbf{X}1}))\,\mathbf{y}$ on mitmemõõtmelise normaaljaotusega (kui $\mathbf{y}$-on normaaljaotusega) siis piisab sõltumatuse näitamiseks vaid sellest, kui suudame näidata võrdust $\text{cov}\left[(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})\mathbf{y}, (\mathbf{I} - \mathbf{P}_{\mathbf{X}1})\mathbf{y}\right] = 0$:

$$\begin{aligned}\text{cov}\left[(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})\mathbf{y}, (\mathbf{I} - \mathbf{P}_{\mathbf{X}1})\mathbf{y}\right] &= (\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})\text{cov}(\mathbf{y},\mathbf{y})(\mathbf{I} - \mathbf{P}_{\mathbf{X}1})^T \\ &= \sigma^2(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})(\mathbf{I} - \mathbf{P}_{\mathbf{X}1})^T \\ &= \sigma^2(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0} - \mathbf{P}_{\mathbf{X}1} + \mathbf{P}_{\mathbf{X}0}\mathbf{P}_{\mathbf{X}1}) \\ &= 0.\end{aligned}$$

Seega võime järeldada, et

$$\frac{\mathbf{y}^T(\mathbf{P}_{\mathbf{X}1} - \mathbf{P}_{\mathbf{X}0})\mathbf{y}/(p_1 - p_0)}{\mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}1})\mathbf{y}/(n - p_1)} \overset{H_0}{\sim} F_{p_1-p_0;n-p_1}.$$

Muuseas, mõnes kontekstis võib olla kasu teadmisest, et

$$\begin{aligned}\mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} &= \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0} - (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1}))\mathbf{y} \\ &= \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} - \mathbf{y}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y} \\ &= SSE_0 - SSE_1\end{aligned}$$

ehk prognooside erinevuste ruutude summa on sama, mis lihtsama mudeli jääkide ruutude summa (SSE_0) miinus keerukama mudeli jääkide ruutude summa (SSE_1).