

4.2 F-testi geomeetrisest interpretatsioonist

Vaatlusvektor $\mathbf{y}$ „elab“ vektorruumis $\mathbb{R}^n$. Jagame selle vektorruumi kolmeks (ortogonaalseks) osaks:

$$\mathbb{R}^n = \mathcal{C}(\mathbf{X}_1)^\perp \oplus \mathcal{C}(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) \oplus \mathcal{C}(\mathbf{P}_{\mathbf{X}_0}).$$

Pane tähele: $\mathcal{C}(\mathbf{P}_{\mathbf{X}_1}) = \mathcal{C}(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) \oplus \mathcal{C}(\mathbf{P}_{\mathbf{X}_0})$.

F-test kontrollib, kas vaatlusvektori vari (matemaatilises keeles projektatsioon) ruumis $\mathcal{C}(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})$ on sama hüplev kui ruumis $\mathcal{C}(\mathbf{X}_1)^\perp$ (kus kogu varieeruvus on meile teadaolevalt põhjustatud vaid juhusest). Kui kahtlusaluses osas vaatlusvektori varieeruvus on samasugune nagu taustavarieeruvus, siis pole meil otseselt vajadust oma lihtsama mudeli mudelimaatriksi poolt genereeritavat vektorruumi $\mathcal{C}(\mathbf{X}_0)$ täiendada.

Näide.

Tehakse seitse mõõtmist. Kolme esimese mõõtmise ajal oleme kindlad, et signaali pole — kõik mida näeme, on müra: $Y_i = \varepsilon_i, i = 1..3$. Kaks viimast mõõtmist sisaldavad kindlalt signaali, $Y_i = \mu_i + \varepsilon, i = 6,7$. Aga vahepealsed mõõtmised — nende puhul pole me kindlad. Võib olla, et vaatluste 4 ja 5 korral kehtib lihtsam mudel, $Y_i = \varepsilon_i$, aga võib olla algas signaali edastamine varem, ja õigem oleks keerukam variant — $Y_i = \mu_i + \varepsilon$. Antud juhul on lihtsam mudel erijuht keerukamast.

$$\begin{array}{ccccccc} \begin{pmatrix} 2 \\ -10 \\ 5 \\ -4 \\ 8 \\ 89 \\ 60 \end{pmatrix} & = & \begin{pmatrix} 2 \\ -10 \\ 5 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} & + & \begin{pmatrix} 0 \\ 0 \\ 0 \\ -4 \\ 8 \\ 0 \\ 0 \end{pmatrix} & + & \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 89 \\ 60 \end{pmatrix} \\ \mathbf{y} & & (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y} & & (\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} & & \mathbf{P}_{\mathbf{X}_0}\mathbf{y} \\ \text{rank} = 6 & & \text{rank} = 3 & & \text{rank} = 2 & & \text{rank} = 2 \\ \text{vaatlusvektor} & & \text{kindlasti müra} & & \text{ebaselge} & & \text{kindlasti midagi} \end{array}$$

Hindame jääkide dispersiooni kahel viisil.

Dispersioonihinnangu $\mathbf{y}^T(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}/\text{rank}(\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) = ((-4)^2 + 8^2)/2$ puhul jagame jääkide ruutude summat kahega, sest (sisuliselt) kasutame esialgsetest vaatlustest vaid kahte.

Kui hindame jääkide dispersiooni valemiga $\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{X}_1})\mathbf{y}/\text{rank}(\mathbf{I}-\mathbf{P}_{\mathbf{X}_1}) = (2^2 + (-10)^2 + 5^2)/3$ siis jagame ruutude summa läbi 3-ga, sest sisuliselt kasutame dispersiooni hindamiseks vaid kolme vaatlust — teisi ruumisuundi me ignoreerime.

Kahe mudeli võrdlemiseks kasutatava *F*-statistiku väärtuseks saame $F = \frac{40}{43} \approx 0,93$ mis annab *F*-testi olulisustõenäosuseks 0,48... Seega jääb nullhüpotees kummutamata — on võimalik, et vaatluste 4 ja 5 ajal signaali edastamist ei toimunud.

4.3 *F*-test hüpoteeside kontrollimiseks mudeli parameetrite kohta

Hinnatava parameeterfunktsiooni $\boldsymbol{\lambda}^T \boldsymbol{\beta}$ hinnangu jaotuseks (kui vaatlusvektor $\mathbf{y}$ on normaaljaotusega; kehtib ligikaudu ka siis, kui on tegemist suure valimiga...) on normaaljaotus :

$$\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\lambda}^T \boldsymbol{\beta}; \boldsymbol{\lambda}^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{\lambda} \sigma^2).$$

Nullhüpoteesi $H_0 : \boldsymbol{\lambda}^T \boldsymbol{\beta} = \boldsymbol{\lambda}^T \boldsymbol{\beta}_0$ kehtides on seega

$$\frac{\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T \boldsymbol{\beta}_0}{\sqrt{\boldsymbol{\lambda}^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{\lambda} \sigma^2}} \sim N(0; 1).$$

ehk

$$\frac{(\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T \boldsymbol{\beta}_0)^T (\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T \boldsymbol{\beta}_0)}{\boldsymbol{\lambda}^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{\lambda} \sigma^2} \sim \chi_{\text{df}=1}^2.$$

Kuna $\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}}$ oli sõltumatu MSE'st (vaata peatükki 3), saame:

$$\frac{(\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T \boldsymbol{\beta}_0)^T (\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T \boldsymbol{\beta}_0)}{\boldsymbol{\lambda}^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{\lambda} \cdot MSE} \stackrel{H_0}{\sim} F_{\text{df}_1=1; \text{df}_2=n-\text{rank}(\mathbf{X})}.$$

F-testi kasutav lähenemisviis on mugavam, kui soovime kontrollida mitut küsimust samaaegselt. Näiteks tahame teada, kas samaaegselt kehtivad väited $\boldsymbol{\lambda}_1^T \boldsymbol{\beta} = \boldsymbol{\lambda}_1^T \boldsymbol{\beta}_0$ ja $\boldsymbol{\lambda}_2^T \boldsymbol{\beta} = \boldsymbol{\lambda}_2^T \boldsymbol{\beta}_0$. Sellisel juhul võime moodustada maatriksi $\boldsymbol{\Lambda}^T$ pannes üksteise peale vektorid $\boldsymbol{\lambda}_1^T$ ja $\boldsymbol{\lambda}_2^T$. Nullhüpoteesi kehtides oleks siis

$$\boldsymbol{\Lambda}^T \hat{\boldsymbol{\beta}} - \boldsymbol{\Lambda}^T \boldsymbol{\beta}_0 \stackrel{H_0}{\sim} N(0; \boldsymbol{\Lambda}^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{\Lambda} \sigma^2)$$

4.3. F-TEST HÜPOTEESIDE KONTROLLIMISEKS MUDELI PARAMEEETRITE KOHTA 47

ja

$$(\mathbf{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}\sigma^2)^{-1/2}(\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{\Lambda}^T\boldsymbol{\beta}_0) \sim N(0; \mathbf{I})$$

ning

$$(\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{\Lambda}^T\boldsymbol{\beta}_0)^T (\mathbf{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda}\sigma^2)^{-1} (\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{\Lambda}^T\boldsymbol{\beta}_0) \sim \chi^2_{\text{df}=\text{rank}(\mathbf{\Lambda})}.$$

Seega saaksime meid huvitava hüpoteesipaari testimiseks kasutada järgmist F-testi:

$$\begin{aligned} (\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{\Lambda}^T\boldsymbol{\beta}_0)^T (\mathbf{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{\Lambda} \cdot \text{MSE})^{-1} (\mathbf{\Lambda}^T\hat{\boldsymbol{\beta}} - \mathbf{\Lambda}^T\boldsymbol{\beta}_0) / \text{rank}(\mathbf{\Lambda}) \sim \\ \sim F_{\text{df}_1=\text{rank}(\mathbf{\Lambda}); \text{df}_2=n-\text{rank}(\mathbf{X})}. \end{aligned}$$

4.4 Testi võimsus

Mudelite võrdlemisel F-testi abil eeldasime, et keerukam mudel on kindlasti õige. Lihtsam kahest võrreldavast mudelist võis aga ei pruukinud olla õige. Nullhüpotees väitis, et ka lihtsam mudel on õige. Mis juhtub aga siis, kui lihtsam mudel pole õige? Kas ja millise tõenäosusega suudame F-testi abil tõestada, et lihtsam mudel pole õige? Alljärgnevalt arvutamegi, kui hea on F-testi võime tõestada alternatiivset hüpoteesi ehk milline on F-testi võimsus.

Dramatis personae:

- $\mathbf{M}_0 : \mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \varepsilon$.
Lihtsaim mudelitest. Nullhüpoteesi väide. Siin peatükis osutub valeks valikuks.
- $\mathbf{M}_1 : \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon$.
Rikkam meie poolt kaalutav mudel. Õige mudel (kuid võib sisaldada mittevajalikke parameetreid, osad parameetervektori $\boldsymbol{\beta}_1$ elemendid võivad olla ka nullid).

Meenutame varem tõesatud teoreemi 3.1:

Olgu $\mathbf{y} \sim N(\boldsymbol{\mu}; \mathbf{I})$, ja olgu maatriks $\mathbf{A}$ ortogonaalne projektor (sümmeetriline ja idempotentne). Siis $\mathbf{y}^T \mathbf{A} \mathbf{y} \sim \chi^2(r, ncp)$, kus $r := \text{rank}(\mathbf{A})$, $ncp := \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}$.

Ülalmainitud teoreemist järeldus (vaata näiteks valemit 3.1), et

$$\mathbf{y}^T (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1}) \mathbf{y} / \sigma^2 \sim \chi_{df=n-p_1}^2$$

ja juhul kui kehtib keerukam mudel, $E(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\beta}_1$:

$$\begin{aligned} (\mathbf{y}/\sigma)^T (\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) (\mathbf{y}/\sigma) &\sim \chi_{df=p_1-p_0; ncp=E\mathbf{y}^T (\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) E\mathbf{y}/\sigma^2} \\ \mathbf{y}^T (\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) \mathbf{y} / \sigma^2 &\sim \chi_{df=p_1-p_0; ncp=\boldsymbol{\beta}_1^T \mathbf{X}_1^T (\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) \mathbf{X}_1 \boldsymbol{\beta}_1 / \sigma^2} \end{aligned}$$

Juhul, kui õige on vaid keerukam mudel, siis antud avaldises mittetsentraalsuse parameeter ei muutu nulliks. Selgitame siinkohal veidi lähemalt mittetsentraalsuse parameetri arvutuseeskirja. Esmalt asendame kõik vaatlused oma andmestikus nende keskväärtustega. Seejärel leiame sellise vigadeta andmestiku pealt prognoosid samadele vaatlustele mõlema mudeli abil — nii lihtsama kui ka keerukama (õige) mudeli abil. Saadud prognooside erinevuste ruutude summa jagame jääkide dispersiooniga ja saamegi mittetsentraalsuse parameetri väärtuse.

Definitsioon 4.1 *Olgu*

$$X \sim \chi_{df_1;ncp}^2; \quad Y \sim \chi_{df_2}^2$$

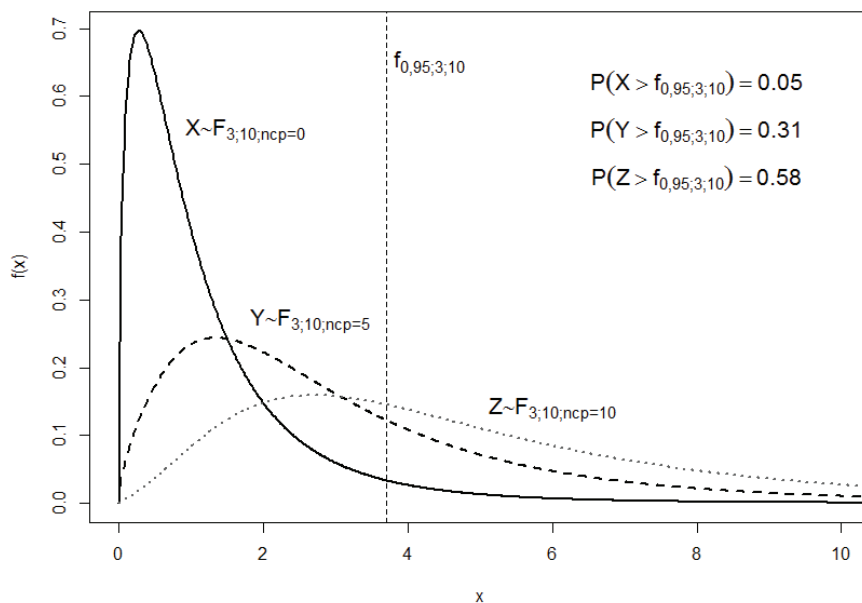
siis juhusliku suuruse

$$Z = \frac{X/df_1}{Y/df_2}$$

jaotuseks on mittetsentraalne F -jaotus vabadusastmete arvudega df_1 , df_2 ja mittetsentraalsuse parameetriga ncp .

Tsentraalse ja mittetsentraalse F -jaotuse tihedusfunktsioonid on toodud ka joonisel 4.1.

Joonis 4.1: Tsentraalne ja mittetsentraalne F -jaotus



Seega tegelik F -statistiku jaotus juhul, kui õige on alternatiivne hüpotees (keerukam mudel) on

$$F \sim F_{df_1=p_1-p_0;df_2=n-p_1;ncp=\beta_1^T \mathbf{X}_1^T (\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) \mathbf{X}_1 \beta_1 / \sigma^2}.$$

Kuna testi võimsus on tõenäosus kummutada nullhüpotees (ehk saada kriitilisest väärtusest suurem teststatistiku väärtus), siis võime testi võimsust leida valemiga

$$1 - F_{df_1=p_1-p_0; df_2=n-p_1; ncp=\beta_1^T \mathbf{X}_1^T (\mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{X}_0}) \mathbf{X}_1 \beta_1 / \sigma^2} (f_{0,95;p_1-p_0;n-p_1}),$$

kus $f_{0,95;p_1-p_0;n-p_1}$ tähistab tsentraalse F-jaotuse $df_1 = p_1 - p_0 (= \text{rank}(\mathbf{X}_1) - \text{rank}(\mathbf{X}_0))$; $df_2 = n - p_1 (= n - \text{rank}(\mathbf{X}_1))$ 0,95-kvantiili.

Näide 4.4 Raha jätkub süvauuringute tegemiseks kümnele saarlasele, kümnele virumaalasele ja veerandsajale tartlasele (tartlaste puhul ei pea me maksma sõidu- ega ööbimiskulusi). Varasemate uuringute põhjal oletame, et uuritava tunnuse standardhälve on ligikaudu $\sigma \approx 25$. Uuritava tunnuse kesk-väärtused on ühe vastava ala tippspetsialisti arvates järgmised: $\mu_{saarlased} = 100$; $\mu_{virulased} = 125$; $\mu_{tartlased} = 108$. Milline on tõenäosus, et suudame F-testi abil tuvastada piirkondlike erinevuste olemasolu?

Lahendus:

Lihtsama mudeli $Y_i = \mu + \varepsilon_i$ prognoos kõigile inimestele (kui inimeste kesk-väärtused oleksid täpselt teada) oleks üldkeskmine,

$$\hat{\mu} = (100 \cdot 10 + 125 \cdot 10 + 108 \cdot 25) / (10 + 10 + 25) = 110.$$

Keerukama mudeli prognoosiks oleks saarlastele saarlaste kesk-väärtus jne. Prognooside erinevuste ruutude summa tuleks

$$(100 - 110)^2 \cdot 10 + (125 - 110)^2 \cdot 10 + (108 - 110)^2 \cdot 25 = 3350$$

ja mittetsentraalsuse parameetriks saaksime $ncp = 3350/25^2 = 5,36$. Testi võimsuse leidmiseks peaksime veel tabelist või arvutist välja uurima F-jaotuse ($df_1 = 3 - 1$, $df_2 = 45 - 3$) kriitilise väärtuse ($f_{0,95;2;42} = 3,22$) ja võimegi leida, millise tõenäosusega kavandatava uuringu abil on võimalik tõestada alternatiivse hüpoteesi paikapidavust (testi võimsus):

$$1 - F_{df_1=2; df_2=42; ncp=5,36}(3,22) = 0,5027 \dots$$

Kui kasutaksime kavandatud valimimahte, siis oleks regionaalsete erinevuste olemasolu tõestamise šansid ligikaudu 1:1.