Peatükk 6

Mitmese võrdluse meetodid

Testides mõne faktortunnuse olulisust (näiteks: kas tunnus *maakond* peaks olema mudelis) võib selguda, et testitav tunnus osutubki statistiliselt oluliseks. Aga peale statistiliselt olulise tulemuse nägemist tekib paratamatult tahtmine teada, millised maakonnad siis teineteisest erinevad. Võiksime muidugi võrrelda kõiki maakondi omavahel — tehes 105 võrdlust. Sedavõrd paljude võrdluste korral tekib aga reaalne oht, et võime näha mõnda statistiliselt olulist tulemust lihtsalt tänu juhusele. Kui iga testi korral nõuaksime, et esimest liiki vea tegemise tõenäosus ei tohi olla suurem kui 0,05 siis 105 testi peale kokku tekib juba väga suur tõenäosus, et mõni nendest testidest eksib ja annab valepositiivse tulemuse (ekslikult kummutab nullhüpoteesi). See võib aga viia situatsioonideni, kus üritame tõlgendada mõnda juhuse tahtel ilmnenud erinevust või peame tõestatuks mõnda erinevust maakondade vahel mida tegelikult ei eksisteerigi.

Võiksime muidugi kasutada mitmeid universaalseid mitmese testimise meetodeid — näiteks Bonferroni või Bonferron-Holmi meetodit. Siiski võiksime endalt küsida — kas lineaarsete mudelite puhul ehk ühes küllaltki spetsiifilises situatsioonis ehk ei teki võimalust leida midagi paremat, midagi täpsemat ja terasemat kui Bonferron(-Holmi) meetod?

Selgub, et teatud juhtudel on tõsti võimalik Bonferron-Holmi meetodit ületada (saavutada suuremat võimsust tegemata siiski kõigi testide peale kokku mõnda esimest liiki viga suurema tõenäosusega kui lubatud). Vaatamegi järgnevalt meetodeid mida sageli kasutatakse lineaarsete mudelite kasutamisel esilekerkivate mitmese testimise probleemide lahendamiseks.

6.1 Scheffé meetod

Bonferrroni meetodi korral peaksime k testi korral kasutama iga üksiktesti tegemisel olulisuse nivood α/k — see tagab, et kõigi testide peale kokku esimest liiki vea tegemise tõenäosus poleks suurem kui α. Aga mis juhtub siis, kui kontrollitavate testide arv läheneks lõpmatusele? Siis muidugi Bonferroni meetodi korral ei õnnestuks mitteühelgi juhul tõestada alternatiivset hüpoteesi — sest olulisuse nivoo läheneks piiramatult nullile. Scheffè meetodit kasutades saaksime aga testida kasvõi lõpmatult palju erinevaid hüpoteese säilitades siiski võime võtta reaalselt vastu alternatiivset hüpoteesi — ja samas tagades, et nende lõpmatult paljude hüpoteeside peale kokku esimest liiki viga liiga sageli ei juhtuks.

Kus võiks tarvis minna kontrollida lõpmatult palju hüpoteese? Näiteks kirjeldame x-tunnuse ja y-tunnuse vahelist seost mingi kõrgema astme polünoomi kasutades. Soovime leida võimalikke käänukohti — kus tunnustevahelist seost kirjeldav funktsioon muutub kasvavast kahanevaks või vastupidi. Iga x-tunnuse väärtuse korral võiksime testida, kas hinnatud funktsiooni tuletis võib olla null. Kus tuletis on null seal funktsioon ei kasva ega kahane — seega võiks olla tegemist käänukohaga. Kui soovime aga märkida näiteks joonisele kõik piirkonnad vahemikus $x = 0 \ldots x = 10$, kus hinnatud funktsiooni tuletis võiks olla null, siis peaksime tegema põhimõtteliselt lõpmatult palju teste. Kui tahaksime samas tagada, et kõigi nende lõpmatult paljude testide peale kokku ei thetaks esimest liiki viga suurema tõenäosusega kui 0,05 (st tahame tagada, et kõik joonisele kantud „kasvupiirkonnad“ ja „kahanemispiirkonnad“ ka tegelikult märgiksid alasid, kus y-tunnuse keskväärtus x-tunnuse väärtuste kasvades vastavalt kas kasvab või kahaneb) siis Bonferroni meetodit kasutades ei õnnestuks meil leida ühtegi piirkonda, kus funktsiooni kasvusuund oleks teada (sest lõpmatu arvu testide tõttu langeb üksiktesti tegemisel kasutatav olulisuse nivoo nulliks). Scheffé meetod aga võib meid hädast välja aidata ja pakkuda siiski võimalust mõnedel juhtudel kindlalt öelda, et siin piirkonnas funktsioon tegelikult kasvab ja siin piirkonnas kahaneb.

6.1.1 t-testi ja mudelite võrdlemiseks mõeldud F-testi võrdlus

Scheffé meetodi mõistmiseks on kasulik esamlt aru saada, et iga hüpoteesi lineaarse mudeli parameerite kohta, mida saab kirja panna kujul $H_0 : \boldsymbol{\lambda}^T\boldsymbol{\beta} = 0$, on võimalik esitada ka kui kahe mudeli võrdlemist (mida saab thea F-testi abil). Näiteks kui vaatame mudelit $y = c_0+c_1x+c_2x^2+\varepsilon$, $\boldsymbol{\beta}^T = (c_0; c_1; c_2)$ siis testides hüpoteesi $H_0 : (0; 0; 1)\boldsymbol{\beta} = 0$ (mis on tegelikult hüpotees $H_0 : c_2 = 0$) võiksime t-testi asemel teha ka F-testi võrdlemaks järgnevat kahte mudelit:

$$\begin{aligned}\mathcal{M}_0 &: y = c_0 + c_1 x + \varepsilon\\ \mathcal{M}_1 &: y = c_0 + c_1 x + c_2 x^2 + \varepsilon.\end{aligned}$$

Veendumaks, et mistahes lineaarset hinnatavat hüpoteesi parameetrite kohta (mille kontrollimiseks sobib t-test) on võimalik asendada mudelite võrdlemiseks mõeldud F-testiga peame aga mõnevõrra vaeva nägema. Kui sinus antud väide küsimusi ei tekita, siis võid kohe suunduda järgmise peatüki juurde. Kui aga kahtled, kas antud väide ikka alati kehtib, siis võid proovida jälgida alltoodud arutlust.

Kui testisime t-testi abil hüpoteesi mõne parameetrite hinnatava lineaarkombinatsiooni kohta

$$H_0 : \ \boldsymbol{\lambda}^T\boldsymbol{\beta} (= \mathbf{v}^T\mathbf{X}\boldsymbol{\beta}) = 0$$

kasutasime varem (vaata peatükki 3) t-statistikut kujul

$$\begin{aligned} t &= \frac{\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}}{\sqrt{\widehat{\mathrm{D}\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}\right)}}} = \frac{\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}}{\sqrt{MSE\cdot\boldsymbol{\lambda}^T\left(X^TX\right)^-\boldsymbol{\lambda}}}\\ &= \frac{\mathbf{v}^T\mathbf{P}_{\mathbf{X}}\mathbf{y}}{\sqrt{MSE\mathbf{v}^T\mathbf{P}_{\mathbf{X}}\mathbf{v}}} \overset{H_0}{\sim} t_{df=n-\mathrm{rank}(\mathbf{X})}.\end{aligned}$$

Samas on lisakitsendus $\mathbf{v}^T\mathbf{X}\boldsymbol{\beta} = 0$ vaadeldav kui nõue, et vaatlusvektor ei tohi paikneda vabalt vektorruumis $\mathcal{C}(\mathbf{X})$, vaid peab paiknema vektorruumi $\mathcal{C}(\mathbf{X})$ mingis kindlas alamruumis. Selliseks alamruumiks on näiteks mudelimaatriksi

$$\mathbf{X}_0 := (\mathbf{I} - \mathbf{P}_{\mathbf{P}_{\mathbf{X}}\mathbf{v}})\mathbf{X} \tag{6.1}$$

veergude poolt moodustatud vektorruum $\mathcal{C}(\mathbf{X}_0)$. Nimelt $\mathcal{C}(\mathbf{X}_0) \subseteq \mathcal{C}(\mathbf{X})$, sest

$$\begin{aligned}(\mathbf{I} - \mathbf{P}_{\mathbf{P}_{\mathbf{X}}\mathbf{v}})\mathbf{X} &= (\mathbf{I} - \mathbf{P}_{\mathbf{P}_{\mathbf{X}}\mathbf{v}})\mathbf{P}_{\mathbf{X}}\mathbf{X}\\ &= (\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}}\mathbf{v}(\mathbf{v}^T\mathbf{P}_{\mathbf{X}}\mathbf{v})^-\mathbf{v}^T\mathbf{P}_{\mathbf{X}})\mathbf{X}\\ &= \mathbf{P}_{\mathbf{X}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}\mathbf{v}(\mathbf{v}^T\mathbf{P}_{\mathbf{X}}\mathbf{v})^-\mathbf{v}^T\mathbf{P}_{\mathbf{X}})\mathbf{X}\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T(\mathbf{I} - \mathbf{P}_{\mathbf{P}_{\mathbf{X}}\mathbf{v}})\mathbf{X}\\ &= \mathbf{X}\cdot\mathbf{M},\end{aligned}$$

ehk $\exists \mathbf{M}$, nii et $(\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X} = \mathbf{X}\mathbf{M}$ ja järelikult $\mathbf{X}_0 := (\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X} \in \mathcal{C}(\mathbf{X})$ ning seega ka $\mathcal{C}(\mathbf{X}_0) \subseteq \mathcal{C}(\mathbf{X})$.

Kui aga kehtib nullhüpotees, $\boldsymbol{\lambda}^T\boldsymbol{\beta}(= \mathbf{v}^T\mathbf{X}\boldsymbol{\beta}) = 0$, siis ka $\mathrm{E}\mathbf{y} \in \mathcal{C}(\mathbf{X}_0)$:

$$\begin{aligned} \mathrm{E}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}_{\mathbf{P_X v}}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X}_0) \end{aligned}$$

sest

$$\begin{aligned} \mathbf{P}_{\mathbf{P_X v}}\mathbf{X}\boldsymbol{\beta} &= \mathbf{P_X v}\left(\mathbf{v}^T\mathbf{P_X}\mathbf{P_X v}\right)^{-}\mathbf{v}^T\mathbf{P_X}\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P_X v}\left(\mathbf{v}^T\mathbf{P_X}\mathbf{P_X v}\right)^{-}\mathbf{v}^T\mathbf{X}\boldsymbol{\beta} \\ &= 0, \end{aligned}$$

kuna eelduse kohaselt $\mathbf{v}^T\mathbf{X}\boldsymbol{\beta} = 0$.

Selgub, et pole vahet, kas testime hüpoteesi $H_0 : \boldsymbol{\lambda}^T\boldsymbol{\beta} = 0$ kahepoolse t-testi abil või võrdleme F-testi abil mudeleid mudelimaatriksitega $\mathbf{X}$ ja $\mathbf{X}_0$ ($\mathbf{X}_0$ on defineeritud valemiga 6.1). Näitame esmalt, et F-statistiku väärtus ja t-testi teststatistiku absoluutväärtus $|t|$ on teinetise kaudu leitavad. Täpsemalt: $F = t^2$. Veendume selles:

$$\begin{aligned} t^2 &= \frac{\mathbf{y}^T\mathbf{P_X}\mathbf{v}\mathbf{v}^T\mathbf{P_X}\mathbf{y}}{MSE \cdot \mathbf{v}^T\mathbf{P_X}\mathbf{v}} \\ &= \frac{\mathbf{y}^T\mathbf{P_X}\mathbf{v}(\mathbf{v}^T\mathbf{P_X}\mathbf{v})^{-}\mathbf{v}^T\mathbf{P_X}\mathbf{y}}{MSE}. \end{aligned}$$

Kirjutame välja ka mudelite võrdlemiseks mõeldud F-statistiku. Paneme esmalt tähele, et

$$\mathbf{P}_{\mathbf{X}_0} = (\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X}\left(((\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X})^T(\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X}\right)^{-}((\mathbf{I}-\mathbf{P}_{\mathbf{P_X v}})\mathbf{X})^T$$

ja seega

$$
\begin{aligned}
F &= \frac{\mathbf{y}^T\left(\mathbf{P_X}-\mathbf{P_{X_0}}\right)\mathbf{y}/1}{MSE} \\
&= \mathbf{y}^T\left(\mathbf{P_X}-\mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^-\mathbf{X}_0^T\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}-(\mathbf{I}-\mathbf{P_{P_Xv}})\mathbf{X}\cdot(\mathbf{X}_0^T\mathbf{X}_0)^-\mathbf{X}_0^T\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}-\mathbf{P_X}(\mathbf{I}-\mathbf{P_{P_Xv}})\mathbf{X}\cdot\ldots\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}-\mathbf{P_X}(\mathbf{I}-\mathbf{P_{P_Xv}})(\mathbf{I}-\mathbf{P_{P_Xv}})\mathbf{X}\cdot\ldots\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T(\mathbf{I}-\mathbf{P_{Xv}})\mathbf{P_{X_0}}\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}_0^T\mathbf{P_{X_0}})\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T(\mathbf{I}-\mathbf{P_{P_Xv}})\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}-\mathbf{P_X}(\mathbf{I}-\mathbf{P_{P_Xv}})\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}\left(\mathbf{I}-\mathbf{I}+\mathbf{P_{P_Xv}}\right)\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\left(\mathbf{P_X}\mathbf{P_{P_Xv}}\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\mathbf{P_X}\left(\mathbf{P_X}\mathbf{v}\left(\mathbf{v}^T\mathbf{P_X}\mathbf{v}\right)^-\mathbf{v}^T\mathbf{P_X}\right)\mathbf{y}/MSE \\
&= \mathbf{y}^T\mathbf{P_X}\mathbf{v}\left(\mathbf{v}^T\mathbf{P_X}\mathbf{v}\right)^-\mathbf{v}^T\mathbf{P_X}\mathbf{y}/MSE \\
&= t^2
\end{aligned}
$$

Kuna vaadeldava F-statistiku ja t-statistiku ruudu väärtused on alati samad (ka nullhüpoteesi kehtides), siis on mõelma teststatistiku jaotused nullhüpoteesi kehtides samasugused ja seega on mõlema testi kriitilised väärtused teineteise kaudu leitavad:

$P(F > q) = \alpha \Rightarrow P(t^2 < q) = 1-\alpha \Rightarrow P(-\sqrt{q} < t < \sqrt{q}) = 1-\alpha$. Teades, et t-jaotus on nullpunkti suhtes sümmeetriline järeldub viimasest aga, et $\sqrt{f_{1-\alpha;1;df}} = t_{1-\alpha/2;df} = -t_{\alpha/2;df}$ (kui $\alpha < 0{,}5$) ning seega peavad ka mõlema testi otsused olema alati kooskõlas: kui t-test võtab vastu alternatiivse hüpoteesi siis teeks seda ka temale vastav F-test ja vastupidi.

Kokkuvõttes: kahepoolne t-test annab alati samasuguse tulemuse, kui F-test mis testib lihtsustatud mudelimaatriksi $\mathbf{X}_0$ sobivust (kusjuures $p_0 = p-1$).

6.1.2 Scheffé meetod

Vaatame mudelit $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, kus

$$\mathbf{X} = (\mathbf{X}_0|\mathbf{X}_1), \text{ ja } \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \end{pmatrix},$$

ehk teisisõnu: $\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$. Testime, kas lihtsam mudel $\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0$ on sama hea kui keerukam. Teeme testi, kontrollime, kas

$$\frac{\mathbf{y}^T(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}/(p-p_0)}{MSE} \sim F_{p-p_0;n-p},$$

kus $p := \operatorname{rank}(\mathbf{X})$ ja $p_0 := \operatorname{rank}(\mathbf{X}_0)$.

Kui teostatud test lükkab ümber nullhüpoteesi, kerkib esile küsimus: milles ikkagi seisneb erinevus? Millised faktori nivood on teineteisest erinevad? Millised vaid parameetervektorit $\boldsymbol{\beta}_1$-te puudutavad hinnatavad lineaarkombinatsioonid (kontrastid) $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ on nullist erinevad?

Viimane väide vajab veidi täpsustustamist: kuna puudutab hinnatav parameeterfunktsioon $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ ainult parameetervektori osa $\boldsymbol{\beta}_1$-te? Siis, kui $\boldsymbol{\lambda}^T\boldsymbol{\beta} = \mathbf{v}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{v}^T\mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{v}^T\mathbf{X}_1\boldsymbol{\beta}_1 = \mathbf{v}^T\mathbf{X}_1\boldsymbol{\beta}_1$ ehk kui $\mathbf{v}^T\mathbf{X}_0 = 0$. Alternatiivselt kirja pandult: $\mathbf{P_X}\mathbf{v} \subset \mathcal{C}(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}0})$ (sest kui $\mathbf{v}^T\mathbf{X}_0 = 0$ siis ka $\mathbf{P}_{\mathbf{X}0}\mathbf{v} = 0$ ja $(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}0})\mathbf{v} = \mathbf{P_X}\mathbf{v}$).

Loomulikult võime erinevaid parameetervektorit $\boldsymbol{\beta}_1$ puudutavaid lineaarkombinatsioone $\boldsymbol{\lambda}_1^T\boldsymbol{\beta}, \boldsymbol{\lambda}_2^T\boldsymbol{\beta}, \ldots$ testida ka olulisuse nivool α tavaliste t-testide abil (nagu tegime seda peatükis 3), kuid sellisel juhul võiksime kõigi tehtud testide peale kokku teha I-liiki viga märksa suurema tõenäosusega kui α. Kui aga soovime, et kõikmõeldavate parameetervektorit $\boldsymbol{\beta}_1$ puudutavate hinnatavate lineaarkombinatsioonide testimisel ei tehtaks I-liiki viga kõigi testide peale kokku suurema tõenäosusega kui α, siis tasuks testimiseks kasutada Scheffé meetodit.

Scheffé meetodi korral kummutame nullhüpoteesi $H_0 : \boldsymbol{\lambda}_i^T\boldsymbol{\beta} = 0$ siis, kui

$$\frac{\hat{\boldsymbol{\beta}}^T\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i{}^T\hat{\boldsymbol{\beta}}/(p-p_0)}{MSE\boldsymbol{\lambda}_i^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}_i} > f_{1-\alpha;p-p_0;n-p}.$$

Võrdluseks:

kui sooviksime testida vaid ühte hüpoteesi, siis t-statistikut kasutades (t-statistiku ruutu kasutades) jõuaksime järgmise otsustuseeskirjani: võta vastu alternatiivne hüpotees kui

$$\frac{\hat{\boldsymbol{\beta}}^T\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^T\hat{\boldsymbol{\beta}}}{MSE\boldsymbol{\lambda}_i^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}_i} > f_{1-\alpha;1;n-p}.$$

Kui aga kasutaksime F-testi võrdlemaks mudeleid mudelimaatriksitega $\mathbf{X}$ ja $\mathbf{X}_0$ siis võtaksime vastu alternatiivse hüpoteesi (parameetervektori osa $\boldsymbol{\beta}_1$ on siiski vaja) siis, kui

$$\frac{\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}/(p-p_0)}{MSE} > f_{1-\alpha;p-p_0;n-p}.$$

Järgnev teoreem tõestab, et kui mudelite võrdlemiseks mõeldud F-test võtab vastu alternatiivse hüpoteesi, siis leidub ka selline vaid parameetervektorit $\boldsymbol{\beta}_1$ puudutav hinnatav lineaarkombinatsioon $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ mille puhul Scheffé meetod kummutab nulhüpoteesi $H_0 : \boldsymbol{\lambda}^T\boldsymbol{\beta} = 0$. Samuti kehtib vastupidine väide — kui Scheffé meetodi abil õnnestub mingi parameetervektorit $\boldsymbol{\beta}_1$ puudutava hinnatava lineaarkombinatsiooni jaoks nullhüpotees kummutada, siis peab ka F-test nullhüpoteesi kummutama ja rikkama mudeli kasuks otsustama.

Teoreem 6.1

$$\begin{aligned} \frac{\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}/(p-p_0)}{MSE} &> f_{1-\alpha;p-p_0;n-p} && (6.2) \\ \Leftrightarrow & \\ \exists \mathbf{v}, \mathbf{v}^T\mathbf{X}_0 = 0, & \\ \frac{\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\hat{\boldsymbol{\beta}}/(p-p_0)}{MSE\mathbf{v}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{v}} &> f_{1-\alpha;p-p_0;n-p}. && (6.3) \end{aligned}$$

Tõestuse idee.

Tõestamaks, et kehtib väide

$$a > b \Leftrightarrow \exists \mathbf{x},\ f(\mathbf{x}) > c$$

võime näidata, et

1. $a > b \Rightarrow \exists \mathbf{x},\ f(\mathbf{x}) > c$
2. $a \le b \Rightarrow$ ei leidu sellist vektorit $\mathbf{x}$, mis rahuldaks nõuet $f(\mathbf{x}) > c$
   ehk
   $a \le b \Rightarrow \forall \mathbf{x},\ f(\mathbf{x}) \le c$.

Tõestus (1). Näitame, et kui kehtib (6.2), siis tõepoolest eksisteerib vähemalt üks vektor $\mathbf{v}$, nii et kehtib ka võrratus (6.3). Veendume, et üheks võrratust rahuldavaks vektoriks on vektor

$$\mathbf{v} := (\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}.$$

Paneme tähele, et $\mathbf{v}^T\mathbf{X}_0 = 0$ (sest $\mathbf{X}_0 \subseteq \mathcal{C}(\mathbf{X})$ ja seega ka $\mathbf{P_X}\mathbf{X}_0 = \mathbf{X}_0$).

On lihtne veenduda, et $\boldsymbol{\lambda}^T\boldsymbol{\beta} := \mathbf{v}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{v}^T\mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{v}^T\mathbf{X}_1\boldsymbol{\beta}_1$ on hinnatav ja sõltub vaid parameetervektoris $\boldsymbol{\beta}_1$ asuvatest tundmatutest parameetritest.

Paigutades valitud $\mathbf{v}$ väärtuse võrratusse 6.3 ja arvestades, et $\mathbf{P_X}(\mathbf{P_X} - \mathbf{P_{X_0}}) = \mathbf{P_X} - \mathbf{P_{X_0}}$, saame

$$\frac{\hat{\boldsymbol{\beta}}^T \mathbf{X}^T (\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{X}\hat{\boldsymbol{\beta}}/(p-p_0)}{MSE \cdot \mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{P_X}(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}} > f_{1-\alpha;p-p_0;n-p}$$

$$\frac{\mathbf{y}^T\mathbf{P_X}(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{P_X}\mathbf{y}/(p-p_0)}{MSE \cdot \mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}} > f_{1-\alpha;p-p_0;n-p}.$$

$$\frac{\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}/(p-p_0)}{MSE \cdot \mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}} > f_{1-\alpha;p-p_0;n-p}.$$

$$\frac{\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}/(p-p_0)}{MSE} > f_{1-\alpha;p-p_0;n-p}.$$

Viimane avaldis on aga tõene tehtud eelduse (6.2) kohaselt.

Nüüd võtame ette tõestuse teise poole (2) ja näitame, et

$$\begin{aligned}
\frac{\mathbf{y}^T(\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{y}/(p-p_0)}{MSE} &\leq f_{1-\alpha;p-p_0;n-p} \\
\Rightarrow & \\
\forall \mathbf{v}, \mathbf{v}^T\mathbf{X}_0 = 0 : & \\
\frac{\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\hat{\boldsymbol{\beta}}/(p-p_0)}{MSE\mathbf{v}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{v}} &\leq f_{1-\alpha;p-p_0;n-p}.
\end{aligned}$$

Paneme esmalt tähele, et

$$\begin{aligned}
&\frac{\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\hat{\boldsymbol{\beta}}/(p-p_0)}{\mathbf{v}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{v}} = \\
&= \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{v}(\mathbf{v}^T\mathbf{P_X}\mathbf{v})^-\mathbf{v}^T\mathbf{X}\hat{\boldsymbol{\beta}}/(p-p_0) \\
&= \mathbf{y}^T\mathbf{P_X}\mathbf{v}((\mathbf{P_X}\mathbf{v})^T(\mathbf{P_X}\mathbf{v}))^-(\mathbf{P_X}\mathbf{v})^T\mathbf{y}/(p-p_0) \\
&= \mathbf{y}^T\mathbf{P}_{\mathbf{P_X v}}\mathbf{y}/(p-p_0)
\end{aligned}$$

Aga $\mathbf{P}_{\mathbf{P_X v}}$ on ortogonaalne projektor ühemõõtmelisse alamruumi $\mathcal{C}(\mathbf{P_X v})$, mis on omakorda ruumi $\mathcal{C}(\mathbf{P_X} - \mathbf{P_{X_0}})$ alamruum, $\mathcal{C}(\mathbf{P_X v}) \subseteq \mathcal{C}(\mathbf{P_X} - \mathbf{P_{X_0}})$. Selles veendumiseks märka, et $\mathbf{P}_{\mathbf{P_X v}} = (\mathbf{P_X} - \mathbf{P_{X_0}})\mathbf{P}_{\mathbf{P_X v}}$ ja järelikult on kõik $\mathbf{P}_{\mathbf{P_X v}}$ veerud esitatavad maatriksi $\mathbf{P_X} - \mathbf{P_{X_0}}$ veergude lineaarkombinatsioonide kaudu.

Aga vaesema mudeli (vähemaid tunnuseid kasutava) mudeli prognooside ruutude summa $\mathbf{y}^T\mathbf{P}_{\mathbf{P_X v}}\mathbf{y}$ on aga alati väiksem (või samasuur) kui rikkama mudeli prognooside ruutude summa $\mathbf{y}^T(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}$.

Vahepala
Veendume, et vaesema mudeli prognooside dispersioon (või prognooside ruutude summa) on alati väiksem (või samasuur) kui rikkama mudeli prognooside dispersioon (prognooside ruutude summa).

1. Olgu meil mudel $\mathbf{X} = (\mathbf{X}_0|\mathbf{X}_1)$. Siis $\mathbf{P_X} = (\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0}) + \mathbf{P}_{\mathbf{X}_0}$ ja

$$\begin{aligned} \mathbf{y}^T\mathbf{P_X}\mathbf{y} &= \mathbf{y}^T(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} + \mathbf{y}^T\mathbf{P}_{\mathbf{X}_0}\mathbf{y} \\ &\geq \mathbf{y}^T\mathbf{P}_{\mathbf{X}_0}\mathbf{y}, \end{aligned}$$

sest maatriks $(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0})$ on mittenegatiivselt määratud (sest ta on projektor):

$$\begin{aligned} (\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0}) \cdot (\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0}) &= \mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0} - \mathbf{P}_{\mathbf{X}_0}\mathbf{P_X} + \mathbf{P}_{\mathbf{X}_0} \\ &= \mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0}, \end{aligned}$$

sest

$$\begin{aligned} \mathbf{P}_{\mathbf{X}_0}\mathbf{P_X} &= (\mathbf{P_X}\mathbf{P}_{\mathbf{X}_0})^T \\ &= \mathbf{P}_{\mathbf{X}_0}^T = \mathbf{P}_{\mathbf{X}_0}. \end{aligned}$$

Seega lihtsama mudeli prognooside ruutude summa on alati samasuur või väiksem, kui keerukama mudeli prognooside ruutude summa (ka siis, kui mõlemad mudelid on õiged!).

2. $\mathbf{P}_{(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0})} = \mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0}$.

Vahepala lõpp.

Seega $\mathbf{y}^T(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} \geq \mathbf{y}^T\mathbf{P}_{\mathbf{P_X v}}\mathbf{y}$ ja järelikult

$$\frac{\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{v}\mathbf{v}^T\mathbf{X}\hat{\boldsymbol{\beta}}/(p-p_0)}{MSE\mathbf{v}^T\mathbf{P_X}\mathbf{v}} \leq \frac{\mathbf{y}^T(\mathbf{P_X} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}/(p-p_0)}{MSE}$$

Viimane on aga tehtud eelduse tõttu väiksem kui $f_{1-\alpha;p-p_0;n-p}$.

□

Tõestatud teoreemist järeldub, et tõenäosus teha ühe või enama Scheffé meetodil kontrollitava hüpoteesi puhul I-liiki viga ei ole suurem kui α. Kui tegelikult kehtib nullhüpotees (näiteks faktortunnus pole oluline), siis F-test eksib ja otsustab keerukama mudeli kasuks tõenäosusega α, tõenäosusega

$1-\alpha$ jääme nullhüpoteesi juurde. Aga jäädes nullhüpoteesi juurde jääme nullhüpoteesi juurde ka kõikide vaid seda faktortunnust puudutavate hinnatavate parameeterfunktsioonide Scheffé meetodil testimisel.

NB! Kontrollides vaid lõplikku arvu hüpoteese Scheffé meetodil on tegelik I-liiki vea tegemise tõenäosus enamasti märkimisväärselt madalam kui α, ehk sellisel juhul on tegemist üleliia konservatiivse meetodiga.

Näide. Vaatame lihtsat dispersioonanalüüsi mudelit:

$$y_{kj} = \mu + \alpha_k + \varepsilon_{kj},$$

kus $k = 1{,}2{,}3{,}4$ ja $j = 1, \ldots, N$. Soovime testida järgmiseid hüpoteese:

$$\begin{aligned} \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 &= 0 \quad (\boldsymbol{\lambda}_1^T \boldsymbol{\beta} = 0) \\ \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 &= 0 \quad (\boldsymbol{\lambda}_2^T \boldsymbol{\beta} = 0) \\ \alpha_1 - \alpha_4 &= 0 \quad (\boldsymbol{\lambda}_3^T \boldsymbol{\beta} = 0) \end{aligned}$$

Variant 1: Kui $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ siis kõik testitavad hüpoteesid kehtivad. Seega võime püstitatud hüpoteeside paikapidavust Scheffé meetodil kontrollida järgmiste testide abil:

$$\frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^T \hat{\boldsymbol{\beta}}/3}{MSE \boldsymbol{\lambda}_i^T (X^T X)^- \boldsymbol{\lambda}_i} > f_{1-\alpha;3;4N-4}.$$

Veidi võimsama testi saaksime, kui märkaksime, et $\boldsymbol{\lambda}_3 = (\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)/2$. Seega esitavad vaadeldud kolm hüpoteesi lisanõudeid vaid vektorruumi $\mathcal{C}(\mathbf{X})$ kahe ruumimõõtme kohta. Seega võiksime ka vaadeldud hüpoteese kontrollides võtta vastu alternatiivse hüpoteesi siis, kui:

$$\frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^T \hat{\boldsymbol{\beta}}/2}{MSE \boldsymbol{\lambda}_i^T (X^T X)^- \boldsymbol{\lambda}_i} > f_{1-\alpha;2;4N-4}.$$

Ka taoline otsustuskriteerium tagab, et I-liiki viga kõigi vaadeldavate hüpoteeside peale kokku ei tehta suurema tõenäosusega kui α.

6.1.3 Usalduspiirid Scheffé meetodil

Soovime leida samaaegseid $(1-\alpha)$ usalduspiire kõigile parameetervektorit β_1-te puudutavaile hinnatavaile lineaarkombinatsioonidele. Teisisõnu öeldes

— soovime et tõenäosusega $(1-\alpha)$ oleksid õiged (sisaldaksid tegelikku parameetrite lineaarkombinatsiooni väärtust) kõik leitud usalduspiirid.

Samaaegsete usalduspiiride konstrueerimiseks vaatame vektorit $\mathbf{y}_* := \mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1$. Võime vektorit $\mathbf{y}_*$ käsitleda kui vaatlusvektorit ja testida F-testi abil mudeleid mudelimaatriksitega $\mathbf{X}$ ja $\mathbf{X}_0$. Kuna vaatlusvektori $\mathbf{y}_*$ jaoks sobib ka lihtsam mudel, siis

$$\frac{\mathbf{y}_*{}^T\left(\mathbf{P}_\mathbf{X} - \mathbf{P}_{\mathbf{X}_0}\right)\mathbf{y}_*/(p-p_0)}{MSE} \sim F_{p-p_0;n-p}.$$

Seega

$$P\left(\frac{\mathbf{y}_*{}^T\left(\mathbf{P}_\mathbf{X} - \mathbf{P}_{\mathbf{X}_0}\right)\mathbf{y}_*/(p-p_0)}{MSE} \leq f_{1-\alpha}\right) = 1-\alpha,$$

kus $f_{1-\alpha}$ tähistab $F_{p-p_0;n-p}$-jaotuse $(1-\alpha)$-kvantiili.

Kasutades teoreemi 6.1 ja tähistades $\hat{\boldsymbol{\beta}}_* = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}_*$ saame, et

$$P\left(\bigcap_{\{\boldsymbol{\lambda}|\boldsymbol{\lambda}^T=\mathbf{v}^T\mathbf{X};\mathbf{v}^T\mathbf{X}_0=0\}}\left\{\frac{\hat{\boldsymbol{\beta}}_*^T\boldsymbol{\lambda}\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_*/(p-p_0)}{MSE\,\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\boldsymbol{\lambda}} \leq f_{1-\alpha}\right\}\right) = 1-\alpha.$$

Märkame, et

$$\begin{aligned}
\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_* &= \mathbf{v}^T\mathbf{X}\hat{\boldsymbol{\beta}}_* \\
&= \mathbf{v}^T\mathbf{P}_\mathbf{X}\mathbf{y}_* \\
&= \mathbf{v}^T\mathbf{P}_\mathbf{X}\mathbf{y} - \mathbf{v}^T\mathbf{P}_\mathbf{X}\mathbf{X}_1\boldsymbol{\beta}_1 \\
&= \boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T\boldsymbol{\beta}
\end{aligned}$$

ja seega

$$P\left(\bigcap_{\{\boldsymbol{\lambda}|\boldsymbol{\lambda}^T=\mathbf{v}^T\mathbf{X};\mathbf{v}^T\mathbf{X}_0=0\}}\left\{\frac{|\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T\boldsymbol{\beta}|^2/(p-p_0)}{MSE\,\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^-\lambda} \leq f_{1-\alpha}\right\}\right) = 1-\alpha.$$

Edasi võime liikuda lihtsate algebraliste teisenduste abil:

$$P\left(\bigcap_{\{\boldsymbol{\lambda}|\boldsymbol{\lambda}^T=\mathbf{v}^T\mathbf{X};\mathbf{v}^T\mathbf{X}_0=0\}}\left\{|\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}-\boldsymbol{\lambda}^T\boldsymbol{\beta}|^2\leq(p-p_0)\widehat{\mathrm{D}\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}\right)}f_{1-\alpha}\right\}\right)=1-\alpha.$$

$$P\left(\bigcap_{\{\boldsymbol{\lambda}|\boldsymbol{\lambda}^T=\mathbf{v}^T\mathbf{X};\mathbf{v}^T\mathbf{X}_0=0\}}\left\{|\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}-\boldsymbol{\lambda}^T\boldsymbol{\beta}|\leq\sqrt{(p-p_0)\widehat{\mathrm{D}\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}\right)}f_{1-\alpha}}\right\}\right)=1-\alpha.$$

$$P\left(\bigcap_{\{\boldsymbol{\lambda}|\boldsymbol{\lambda}^T=\mathbf{v}^T\mathbf{X};\mathbf{v}^T\mathbf{X}_0=0\}}\left\{\lambda^T\hat{\boldsymbol{\beta}}-\sqrt{(p-p_0)\widehat{\mathrm{D}\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}\right)}f_{1-\alpha}}\leq\right.\right.$$
$$\left.\left.\leq\boldsymbol{\lambda}^T\boldsymbol{\beta}\leq\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}+\sqrt{(p-p_0)\widehat{\mathrm{D}\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}\right)}f_{1-\alpha}}\right\}\right)=1-\alpha.$$

Saadud tulemust võib kasutada näiteks regressioonikõverale usaldusriba konstrueerimiseks. Kirjeldagu näiteks vaatlusandmeid järgmine regressioonmudel:

$$y_i=\beta_0+\beta_1x_i+\beta_2{x_i}^2+\varepsilon.$$

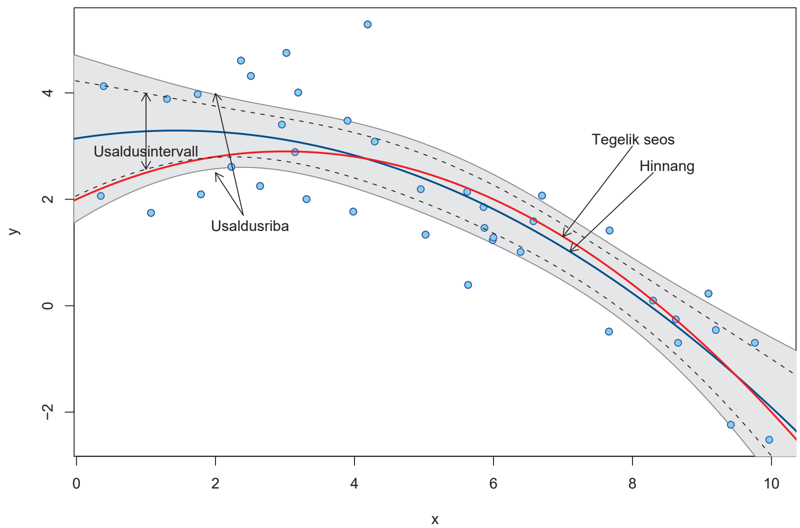
Eeldame, et parameetrite hindamiseks kasutatavas andmestikus leidub vähemalt 3 erinevat tunnuse X väärtust. Sellisel juhul on parameetervektor $\boldsymbol{\beta}^T=(\beta_0;\beta_1;\beta_2)^T$ hinnatav ja $p(=\mathrm{rank}(\mathbf{X}))=3$.

Soovime hinnatud regressioonijoont kujutaval joonisel näidata ka 0,95-usaldusriba — riba, kuhu vahele tegelik regressioonikõver jääb tõenäosusega 0,95 (keskmiselt 95% valimite korral ei kaldu tegelik seost iseloomustav regresioonijoon kasvõi korrakski — mitte ühegi x-i väärtuse korral — joonisele kantud usaldusribast välja).

Usaldusriba konstrueerimiseks tahame leida usaldusintervalle parameetrite lineaarkombinatsioonidele kujul $\beta_0+\beta_1x+\beta_2x^2$. Sealjuures soovime, et tõenäosusega 0,95 saaksime valimi, kus kõigi vaadeldud lineaarkombinatsioonide peale kokku — ükskõik kui palju x-i väärtuseid me ka joonise tegemiseks läbi ei vaataks — ei esineks ühtegi viga (alati kuuluks tegelik parameetrite lineaarkombinatsiooni väärtus väljapakutud vahemikku). Paneme tähele, et antud juhul hõlmavad meid huvitavad lineaarkombinatsioonid kõiki parameetervektori elemente, seega $p_0=0$. Soovitud usaldusvahemikud on seega kujul

$$\hat{\beta}_0+\hat{\beta}_1x+\hat{\beta}_2x^2\pm\sqrt{3\widehat{\mathrm{D}}\left(\hat{\beta}_0+\hat{\beta}_1x+\hat{\beta}_2x^2\right)f_{0,95}},$$

Joonis 6.1: Usaldusriba Scheffe meetodil



kus $f_{0,95}$ on $F_{3;n-3}$-jaotuse 0,95-kvantiil. Saadud usaldusriba iseloomustab joonis 6.1, kus on ära toodud nii Scheffé meetodil konstrueeritud 0,95-usaldusriba kui ka iga x-i väärtuse korral leitud „tavalised“ nn punktiviisilised 0,95-usaldusintervallid.