Föreläsning 9: Linjär regression – del II

Johan Thim (johan.thim@liu.se)

29 september 2018

"No tears, please. It's a waste of good suffering."
–Pinhead

Vi fixerar en vektor $\boldsymbol{u}^T = (1 \ u_1 \ u_2 \ \cdots \ u_k)$, där u_i kommer vara värdet på x_j i den punkt vi kommer betrakta. Vi är alltså intresserade av vad modellen har att säga vid en fixerad punkt där vi inte gjort någon mätning. Vi betraktar Y_0 definierad av

$$Y_0 = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_k u_k + \epsilon_0 = \boldsymbol{u}^T\boldsymbol{\beta} + \epsilon_0.$$

Vi antar att $\epsilon_0 \sim N(0,\sigma^2)$ är oberoende av $\boldsymbol{\epsilon}$. Vi definierar

$$\mu_0 = E(Y_0) = \boldsymbol{u}^T\boldsymbol{\beta}.$$

2 Konfidensintervall för $E(Y_0)$

En naturlig skattning av μ_0 ges av $\boldsymbol{u}^T\widehat{\boldsymbol{\beta}}$, så vi sätter

$$\widehat{\mu_0} = \boldsymbol{u}^T\widehat{\boldsymbol{\beta}}.$$

Eftersom $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$ blir

$$E(\widehat{\mu_0}) = \boldsymbol{u}^T E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{u}^T\boldsymbol{\beta}$$

och

$$V(\widehat{\mu_0}) = \sigma^2\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}.$$

Då $\widehat{\mu_0}$ är en linjärkombination av normalfördelade variabler gäller att

$$\widehat{\mu_0} \sim N(\boldsymbol{u}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}).$$

Således gäller att

$$\frac{\widehat{\mu_0} - \boldsymbol{u}^T\boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim N(0,1).$$

I vanlig ordningen brukar vi behöva skatta σ^2 och gör det med

$$s^2 = \frac{\mathrm{SS_E}}{n-k-1}, \quad \text{där } S^2 \sim \chi^2(n-k-1).$$

Då gäller (enligt Gossets sats) att

$$\frac{\widehat{\mu_0} - \boldsymbol{u}^T\boldsymbol{\beta}}{S\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(n-k-1).$$

Genom att nyttja denna variabel kan vi ställa upp ett tvåsidigt konfidensintervall för $E(Y_0)$:

$$I_{\mu_0} = \left(\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} - t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}},\ \boldsymbol{u}^T\widehat{\boldsymbol{\beta}} + t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}\right).$$

Intervallet I_{μ_0} beskriver vart uppmätta värden vid $\boldsymbol{u}$ hamnar i snitt, dvs vid många upprepningar med samma $\boldsymbol{u}$ så hamnar vi i intervallet. Det säger inget om vart en enskild mätning hamnar, för det behöver vi prediktionsintervall!

3 Prediktionsintervall för Y_0

Vill vi uppskatta (förutsäga) vad mätvärdet y_0 blir i en viss punkt $\boldsymbol{u}$ ställer vi upp ett **prediktionsintervall**. Eftersom $Y_0 \sim N(\mu_0, \sigma^2)$ och $\widehat{\mu_0} = \boldsymbol{u}^T\widehat{\boldsymbol{\beta}} \sim N(\mu_0, \sigma^2\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u})$ är oberoende gäller det att

$$V(Y_0 - \widehat{\mu_0}) = \sigma^2\left(1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}\right)$$

så

$$Y_0 - \widehat{\mu_0} \sim N\left(0,\ \sigma^2\left(1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}\right)\right).$$

Vi skattar σ^2 med s^2 och nyttjar Gossets sats:

$$\frac{Y_0 - \widehat{\mu_0}}{S\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(n-k-1).$$

Vi kan stänga in denna variabel och lösa ut Y_0:

$$I_{Y_0} = \Bigg(\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} - t_{\alpha/2}(n-k-1)s\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}},$$
$$\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} + t_{\alpha/2}(n-k-1)s\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}\Bigg).$$

4 Konfidens- och prediktionsband

Vid grafisk representation av enkel linjär regression ser man ofta så kallade konfidens- och prediktionsband inritade. Dessa definieras enligt följande.

Konfidensband

Definition. Ett **konfidensband** ges av en funktion g sådan att för varje x gäller att

$$P(|\mu_0(x) - \widehat{\mu_0}(x)| < g(x)) = 1 - \alpha.$$

Ett **simultant konfidensband** uppfyller att

$$P(|\mu_0(x) - \widehat{\mu_0}(x)| < g(x) \text{ för alla } x) = 1 - \alpha.$$

Skillnaden mellan ett simultant band och dess icke-simultana motsvarighet kanske är svår att se, men det simultana bandet uppfyller alltså instängningen med sannolikheten $1-\alpha$ för *alla* x på en gång medan den icke-simultana uppfyller denna sannolikhet för varje x *en i taget*! Likformighet är något det simultana bandet erbjuder. Om vi endast har ett icke-simultant konfidensband och vill titta i två punkter x_1 och x_2 samtidigt är det inte säkert att dessa intervall *samtidigt* uppfyller konfidensgraden $1 - \alpha$. Det är precis samma problem vi sett vi beräkningar av flera konfidensintervall samtidigt tidigare.

Prediktionsband

Definition. Ett **prediktionsband** ges av en funktion h sådan att för varje x gäller att

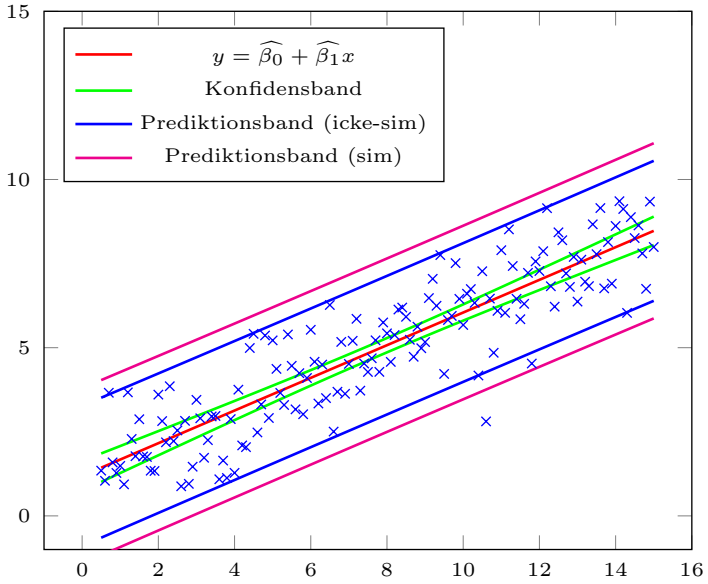
$$P(|y(x) - \widehat{y}(x)| < h(x)) = 1 - \alpha.$$

Ett **simultant prediktionsband** uppfyller att

$$P(|y(x) - \widehat{y}(x)| < h(x) \text{ för alla } x) = 1 - \alpha.$$

Grafiskt kan det se ut enligt nedan. Man ritar ofta i både konfidens- och prediktionsband samtidigt. Notera att konfidensbandet är betydligt smalare än prediktionsbandet.

Konfidensband; $\alpha = 0.05$



5 Residualanalys

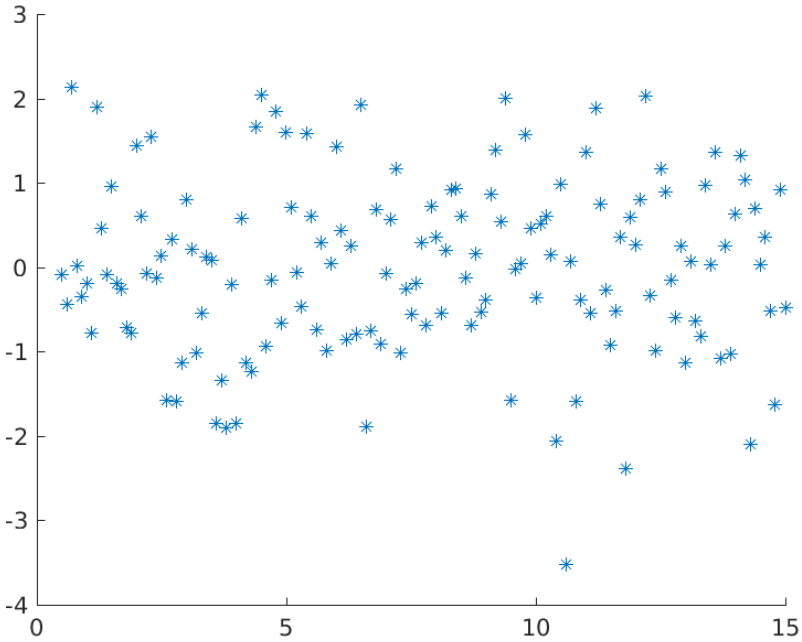
Efter utförd regression har vi skattade y-värden $\widehat{\mu}$ (eller $\widehat{y}$), som används för att beräkna kvadratsumman $\mathrm{SS_E}$ för felen som modellen inte förklarar. Antagandet vi gjort på **residualerna** $e_j = y_j - \widehat{y_j}$ är att dessa är oberoende och normalfördelade med samma varians och väntevärde 0. Detta är något som bör undersökas efter regressionen för att motivera antagandet. I matlab kan vi ta fram residualerna vid regression genom kommandot

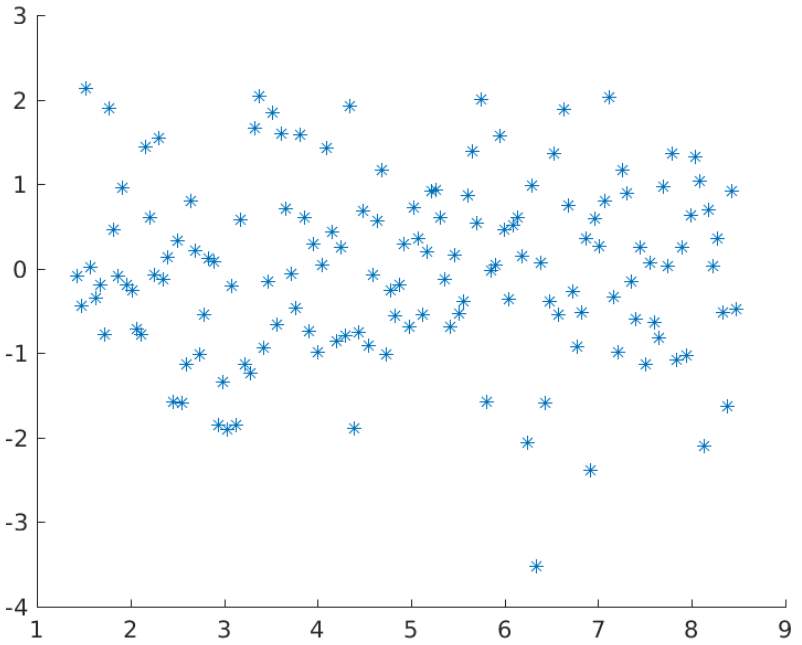
```
>> r = regstats(y, x, 'linear', 'all');
>> res = r.r;
>> yhat = r.yhat;
```

5.1 Residualer vs x eller $\widehat{y}$

Vi kan plotta residualer mot x-värden eller skattade y-värden ($\widehat{y} = \widehat{\mu}$):

```
>> figure; scatter(x, res, '*');
>> figure; scatter(yhat, res, '*');
```



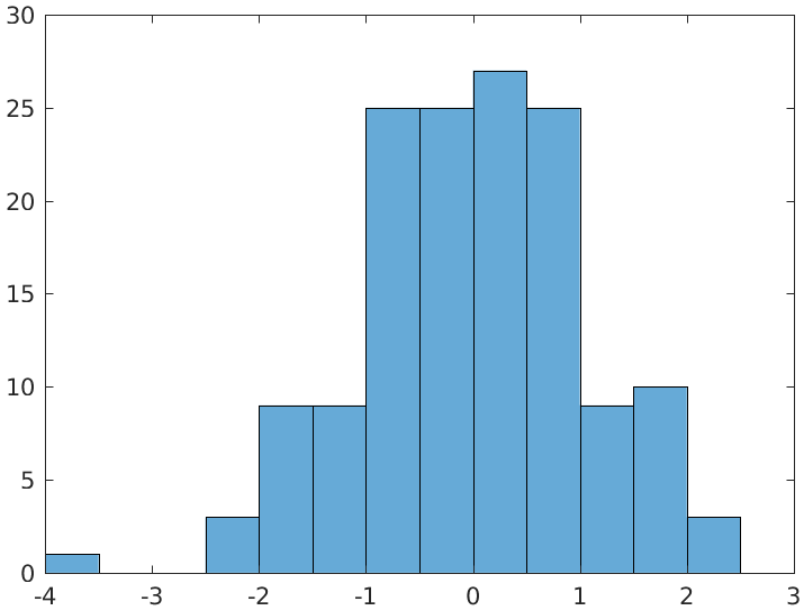


Det är svårt att se något direkt samband. Vilket är bra. Hade vi sett ett tydligt samband hade vi haft problem med modellen. Men mycket mer än så kan vi inte säga från dessa figurer.

5.2 Histogram

Vi kan plotta ett histogram för residualerna:

```
>> figure; histogram(res);
```

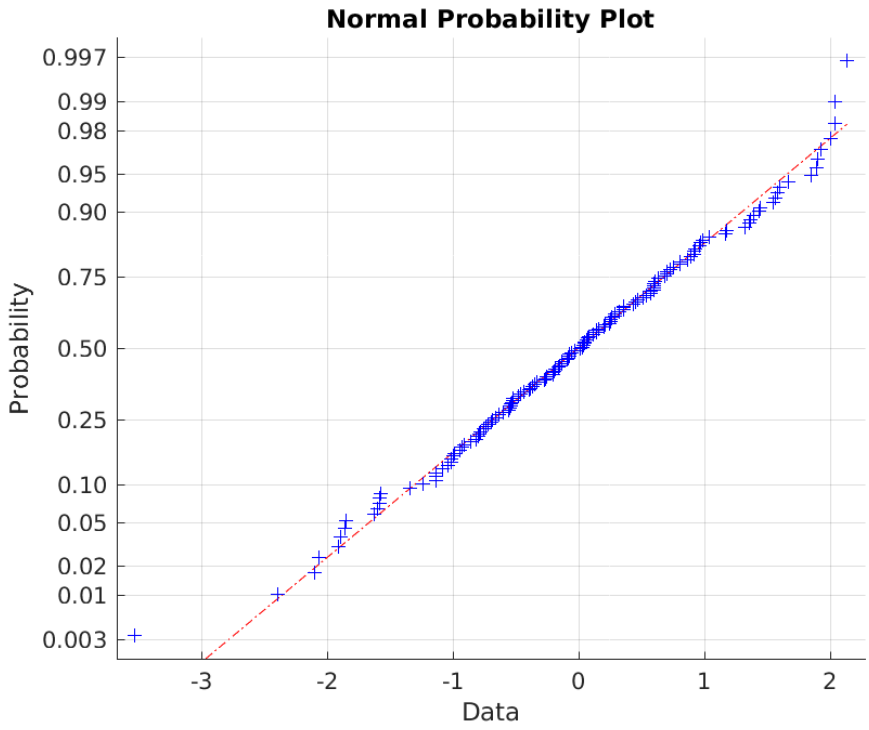


Det ser hyfsat Gaussiskt ut och masscentrum är runt nollan. Inte helt orimligt med normalfördelning.

5.3 Normalplot

Matlab kan även enkelt generera en så kallad normalplot:

```
>> figure; normplot(res);
```



I figuren så skalar alltså y-axeln mot sannolikheter som gäller för normalfördelning (tänk på exempelvis log-skala fungerar). Idealiskt skulle vi endast ha punkter som ligger *på* en linje. Nu finns kanske lite tillstymmelse till så kallad **S-form** på kurvan, men absolut inte på den nivå att vi borde ifrågasätta antagandet kring normalfördelning. Betydligt mer S-lika kurvor skulle accepteras som rimligt normalfördelade.

6 Variabeltransformation

Det vi håller på med kallas linjär regression, men det är inget som hindrar oss att ändå använda linjär struktur för att anpassa ett polynom eller mer generella funktioner till mätdata istället[1]

6.1 Polynomiell regression

Antag att vi vill bestämma ett polynom av grad k som minimerar kvadratfelet. Modellen är att x_j är fixerade tal och att y_j är observationer av

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \cdots + \beta_k x_j^k + \epsilon_j,$$

där $\epsilon_j \sim N(0, \sigma^2)$ är oberoende. Vi löser detta problem med linjär regression genom att låta

$$x_{j1} = x_j, \quad x_{j2} = x_j^2, \quad x_{j3} = x_j^3, \quad \cdots \quad x_{jk} = x_j^k,$$

och sedan betrakta modellen

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + \epsilon_j,$$

där $\epsilon_j \sim N(0, \sigma^2)$ är oberoende.

6.2 Exponentiell regression

Antag att vi har data som verkar vara följa en exponentialkurva. Modellen är att x_j är fixerade tal och att y_j är observationer av

$$Y_j = a \exp(b x_j) \cdot E_j \tag{1}$$

där E_j är **lognormal-fördelade** och oberoende.

Lognormal-fördelning

Definition. Slumpvariabeln X kallas **lognormal-fördelad** med parametrarna μ och σ om

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

Vi skriver $X \sim \text{Lognormal}(\mu, \sigma^2)$.

[1] Helt analogt med vad som gjorts i linjär algebra; tänk polynombaser. Eller Fourieranalys för den delen.

Det följer att $E(X) = \exp(\mu + \sigma^2/2)$ och $V(X) = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$ eftersom

$$E(h(X)) = \int_0^\infty \frac{h(x)}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) dx = \left/ \begin{matrix} y = \ln x \\ dy = \dfrac{dx}{x} \end{matrix} \right/$$

$$= \int_{-\infty}^\infty \frac{h(e^y)}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dx$$

så

$$E(X) = \int_{-\infty}^\infty \frac{e^y}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dx = \left/ \begin{matrix} u\sigma = y - \mu \\ dy = \sigma\, du \end{matrix} \right/$$

$$= \int_{-\infty}^\infty \frac{e^{\mu+\sigma^2/2}}{\sqrt{2\pi}} \exp\left(-\frac{(u-\sigma)^2}{2}\right) dx = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

och på samma sätt blir

$$E(X^2) = \exp\left(2\mu + 2\sigma^2\right).$$

vilket ger

$$V(X) = E(X^2) - E(X)^2 = \left(\exp(\sigma^2) - 1\right)\exp\left(2\mu + \sigma^2\right).$$

Sats. Om $X \sim \text{Lognormal}(\mu, \sigma^2)$ så är $\ln X \sim N(\mu, \sigma^2)$.

Bevis. Låt X vara lognormalfördelad och låt $Y = \ln X$. Eftersom exp är strängt växande gäller att

$$F_Y(y) = P(Y \le y) = P(\ln X \le y) = P(X \le e^y)$$

vilket medför att

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(e^y)\, e^y = \frac{e^y}{e^y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln e^y - \mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Således är $Y = \ln X \sim N(\mu, \sigma^2)$. □

Vi löser nu problemet i (1) med linjär regression genom att logaritmera sambandet:

$$\ln Y_j = \ln a + bx_j + \ln E_j = \beta_0 + \beta_1 x_j + \epsilon_j,$$

där $\epsilon_j \sim N(0, \sigma^2)$ är oberoende. Sen använder vi tekniker vi tagit fram!

7 Val av modell

Så låt oss säga att vi har en mängd mätdata i form av y-värden för en mängd olika värden på variabler $x_1, x_2, \ldots, x_k$. Hur ska vi välja modell? Tillför alla variabler något användbart? Hur jämför vi två olika modeller? Frågorna hopar sig.
Vad man alltid kan göra är att studera skattningen för σ^2. Denna skattning kommer i allmänhet från residualerna och idealiskt skulle dessa i princip vara lika med noll (perfekt lösning). Ett mindre värde på s^2 innebär alltså att modellen förklarar lite mer. Nu kan tilläggas att om man lägger till variabler kommer *alltid* s^2 att bli mindre (varför?), så vi behöver avgöra om skillnaden är signifikant.

(i) Val av variabler. Vilka har vi tillgång till? Vilka kan vi utesluta på grunden att de inte bör ingå i modellen? Är vissa variabler väldigt starkt korrelerade (i så fall kan det vara bättre att bara ta med en)?

(ii) Är sambandet linjärt? Kan det genom någon lämplig transformation skrivas som ett linjärt problem? Om det inte går kommer linjär regression fungera dåligt.

(iii) Vid flera möjliga modellval, hur testar vi om skillnaden mellan modellerna är signifikant? Vi vill inte ta med variabler i onödan.

Vi börjar med att diskutera begreppet **inkapslade modeller** (eller **nästlade**). Modeller där vi i någon mening kan säga den ena är en del av den andra.

8 Inkapslade modeller

Om vi har två modeller att välja mellan med syntesmatriserna X_1 respektive X_2. Vi låter H_1 och H_2 vara respektive hattmatriser, så blir

$$H_1 = X_1(X_1^T X_1)^{-1} X_1^T \quad \text{och} \quad H_2 = X_2(X_2^T X_2)^{-1} X_2^T.$$

Vi låter $\boldsymbol{\beta} \in \mathbf{R}^{k_1+1}$ respektive $\boldsymbol{\beta} \in \mathbf{R}^{k_2+1}$ för de olika modellerna. Dimensionerna för X_1 och X_2 är $n \times (k_1+1)$ respektive $n \times (k_2+1)$.

Inkapslade modeller

Definition. Vi kallar modell 1 för **inkapslad** i modell 2 om

$$V_1 = \{X_1\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbf{R}^{k_1+1}\} \subset \{X_2\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbf{R}^{k_2+1}\} = V_2.$$

Definitionen är lite abstrakt, men vad som säges är att kolonnrummet som spänns upp av X_1 ska vara ett underrum till kolonnrummet som spänns upp av X_2. Exempelvis gäller det att modellen

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

är inkapslad i modellen

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_{k+p} x_{k+p} + \epsilon.$$

Detta följer eftersom de första $k+1$ kolonnerna i X_1 och X_2 är identiska. Detta är i huvudsak vad vi ska använda inkapslade modeller till: att undersöka om det blir signifikant bättre av att lägga till förklaringsvariabler (alternativ att det inte skadar att ta bort förklaringsvariabler).

Sats. Om $V_1 \subset V_2$ gäller att

$$H_1H_2 = H_2H_1 = H_1 \quad \text{och} \quad (I - H_1)(I - H_2) = (I - H_2)(I - H_1) = I - H_2.$$

Vidare gäller att $H_2 - H_1$ är en projektionsmatris med $\text{rank}(H_2 - H_1) = \text{rank}(H_2) - \text{rank}(H_1)$.

Bevis. Eftersom

$$H_1\boldsymbol{y} \in V_1 \subset V_2$$

för alla $\boldsymbol{y}$ följer det att $H_2H_1\boldsymbol{y} = H_1\boldsymbol{y}$. Eftersom H_1 och H_2 är symmetriska så medföljer även att $H_1H_2 = H_1$.
För den andra likheten noterar vi att $V_1 \subset V_2$ implicerar att ortogonalkomplementen uppfyller $V_2^\perp \subset V_1^\perp$. Således blir

$$(I - H_2)\boldsymbol{y} \in V_2^\perp \subset V_1^\perp$$

och

$$(I - H_1)(I - H_2)\boldsymbol{y} = (I - H_2)\boldsymbol{y}.$$

Analogt med ovan följer även att $(I - H_2)(I - H_1) = I - H_2$. Det faktum att $H_2 - H_1$ är en projektionsmatris följer av att den uppenbarligen är symmetrisk och

$$(H_2 - H_1)^2 = H_2^2 - H_2H_1 - H_1H_2 + H_1^2 = H_2 - 2H_1 + H_1 = H_2 - H_1.$$

Således är samtliga egenvärden 0 eller 1 och

$$\operatorname{rank}(H_2 - H_1) = \operatorname{tr}(H_2 - H_1) = \operatorname{tr}(H_2) - \operatorname{tr}(H_1) = \operatorname{rank}(H_2) - \operatorname{rank}(H_1).$$

Den sista likheten på grund av att H_1 och H_2 också är projektionsmatriser. □
Vi kan nu formulera (och bevisa) en variant på regressionsanalysens 2:a huvudsats. Den går att formulera mer generellt, men detta är mer än tillräckligt för våra ändamål.

Regressionsanalysens 2:a huvudsats

Sats. Låt H_1 och H_2 ha rang $k_1 + 1$ respektive $k_2 + 1$. Om $V_1 \subset V_2$ så gäller att:

(i) $\mathrm{SS}_\mathrm{E}^{(2)}$ och $\mathrm{SS}_\mathrm{E}^{(1)} - \mathrm{SS}_\mathrm{E}^{(2)}$ är oberoende;

(ii) $\dfrac{\mathrm{SS}_\mathrm{E}^{(2)}}{\sigma^2} \sim \chi^2(n - k_2 - 1)$;

(iii) samt om $E(\boldsymbol{Y}) = \boldsymbol{\mu_1} = X_1\boldsymbol{\beta_1}$ så är $\dfrac{\mathrm{SS}_\mathrm{E}^{(1)} - \mathrm{SS}_\mathrm{E}^{(2)}}{\sigma^2} \sim \chi^2(k_2 - k_1)$.

Bevis. Vi ser att

$$\mathrm{SS}_\mathrm{E}^{(2)} = \boldsymbol{Y}^T(I - H_2)\boldsymbol{Y}$$

och

$$\mathrm{SS}_\mathrm{E}^{(1)} - \mathrm{SS}_\mathrm{E}^{(2)} = \boldsymbol{Y}^T(I - H_1 - (I - H_2))\boldsymbol{Y} = \boldsymbol{Y}^T(H_2 - H_1)\boldsymbol{Y}.$$

Eftersom

$$(I - H_2)(H_2 - H_1) = H_2 - H_1 - H_2^2 + H_2H_1 = -H_1 + H_1 = 0$$

så kommer $(I - H_2)\boldsymbol{Y}$ och $(H_2 - H_1)\boldsymbol{Y}$ att vara okorrelerade och normalfördelade. Således är dessa variabler oberoende vilket medför punkt (i). Punkt (ii) är identisk med resultatet från regressionsanalysens första huvudsats (se förra föreläsningen). Den sista punkten följer av ett liknande argument som på förra föreläsningen. Först, eftersom $V_1 \subset V_2$, så finns ett $\boldsymbol{\alpha} \in \mathbf{R}^{k_2+1}$ så att $X_2\boldsymbol{\alpha} = X_1\boldsymbol{\beta_1}$. Detta medför att

$$(H_2 - H_1)X_1\boldsymbol{\beta_1} = H_2X_2\boldsymbol{\alpha} - X_1\boldsymbol{\beta_1} = X_2\boldsymbol{\alpha} - X_1\boldsymbol{\beta_1} = X_1\boldsymbol{\beta_1} - X_1\boldsymbol{\beta_1} = 0,$$

så $E((H_2 - H_1)\boldsymbol{Y}) = 0$. Därav följer det att

$$\mathrm{SS}_{\mathrm{E}}^{(1)} - \mathrm{SS}_{\mathrm{E}}^{(2)} = \boldsymbol{\epsilon}(H_2 - H_1)\boldsymbol{\epsilon}^T.$$

Eftersom $H_2 - H_1$ är en projektionsmatris med rang $k_2 - k_1$ och $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$ finns en ON-matris C så att med $\epsilon = C\boldsymbol{Z}$ blir

$$\boldsymbol{\epsilon}(H_2 - H_1)\boldsymbol{\epsilon}^T = \sum_{j=1}^{k_2-k_1} Z_j^2,$$

där $Z_j \sim N(0, \sigma^2)$ är oberoende. Alltså stämmer fördelningen i punkt (iii) eftersom kvadratsumman av oberoende $N(0,1)$-variabler blir χ^2-fördelad med frihetsgraden lika med antalet termer. □

Anmärkning. Om vi inte skulle anta att $E(\boldsymbol{Y}) = \boldsymbol{\mu_1}$ så skulle vi fortfarande erhålla en χ^2-fördelningen, men den blir inte centrerad. Överkurs.

Exempel

Vi har två modeller:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

och

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_{k+p} x_{k+p} + \epsilon.$$

Hur kan man testa om $\beta_{k+1} = \beta_{k+2} = \cdots = \beta_{k+p} = 0$ (dvs om de tillförda variablerna hjälper på en signifikant nivå)?

Lösning. Enligt föregående diskussion är modell 1 inkapslad i modell 2. Låt nollhypotesen ges av

$$H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_{k+p} = 0,$$

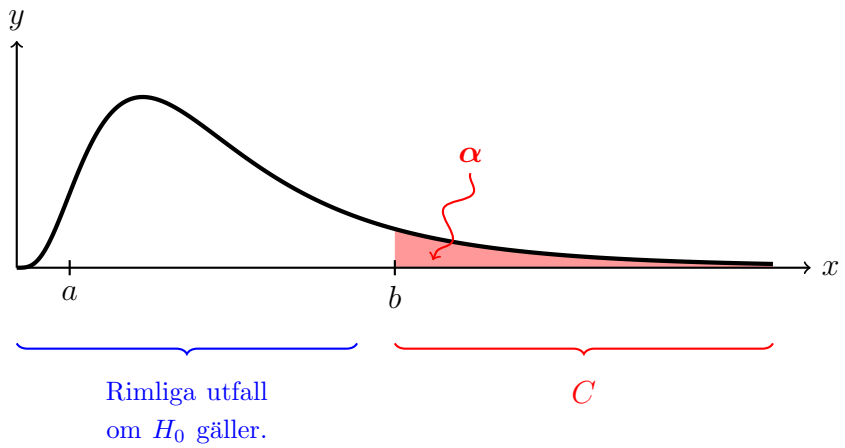
med mothypotesen

$$H_1 : \text{något } \beta_j,\ j = k+1, k+2, \ldots, k+p, \text{ är inte } = 0.$$

Om H_0 är sann så gäller att $\boldsymbol{Y} \sim N(X_1\boldsymbol{\beta_1}, \sigma^2 I)$, så satsen ovan medför direkt att

$$W = \frac{(\mathrm{SS}_{\mathrm{E}}^{(1)} - \mathrm{SS}_{\mathrm{E}}^{(2)})/p}{\mathrm{SS}_{\mathrm{E}}^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1) \quad \text{om } H_0 \text{ är sann}$$

eftersom det är en kvot av oberoende χ^2-fördelade variabler. Om H_0 inte är sann kommer det att göra att W tenderar att bli stor, så vårt kritiska område kommer ges av $C =]c, \infty[$ för något $c > 0$.



9 Stegvis regression

En tänkbar lösning på problemet att hitta en modell som tar med precis de variabler som är signifikanta är givetvis att helt enkelt testa alla kombinationer. Med k möjliga förklaringsvariabler ger det 2^k olika modeller. Vi kan utföra regression för var och en och sedan undersöka vilka variabler som förefaller vara relevanta. Otympligt? Jo, kanske det Så en annan variant är att lägga till en variabel i taget till vi inte ser någon signifikant skillnad längre när vi lägger till fler variabler. Så hur börjar vi?
Den bästa förklaringsvariabeln är alltid den som är starkast korrelerade med y. Detta fenomen följer av exemplet från förra föreläsningen angående enkel linjär regression där vi visade att $\mathrm{SS_E} = (1 - r^2)\sum_{j=1}^{n}(y_j - \overline{y})^2$. Däremot kan vi inte direkt se vilken den näst bästa är utan att utföra en regression. Så processen kommer att se ut enligt följande.

(i) Jämför korrelationen mellan $\boldsymbol{y}$ och de olika x-kolonnerna i X och välj den där r^2 är störst som första förklaringsvariabel.

(ii) Testa och lägg till var och en av resterande variabler en i taget och beräkna $\mathrm{SS_E}$ för varje modell. Välj den variabel som minimerar $\mathrm{SS_E}$. Detta är den nästa bästa förklaringsvariabeln. Lägg till den.

(iii) Testa den nya modellen genom att endera göra ett F-test för att se om den är signifikant bättre eller gör ett t-test för att se om hypotesen $H_0 : \beta_i = 0$ för den tillagda β_i kan förkastas. Om variabeln inte tillför något är vi färdiga. Annars lägg till variabeln i modellen.

(iv) Upprepa steg 2 tills dess att vi inte får någon signifikant skillnad när vi lägger till en ny variabel.


Vi kan *endast* hitta den bästa förklaringsvariabeln genom att studera korrelation mellan y och de olika x_i-variablerna. Eventuell övrig information från exempelvis kovariansmatrisen ger inte nödvändigvis någon information om vad som blir bäst när man väl tagit med den bästa variabeln. Ny analys krävs efter regressionssteget!

10 Kategorier och ”dummy”-variabler

Ibland har man data som är beroende av någon storhet som är binär (eller åtminstone har diskreta nivåer). Till exempel skulle det kunna handla om en modell för åtgång av förbrukningsvaror hos ett café vid stranden. Beroende på om det är sommar eller vinter kanske saker och ting ser helt annorlunda ut. Vi kan då lägga till en variabel i modellen som har värdet 1 vid sommar och 0 när det är vinter. På det sättet kan vi ta med all data i en och samma modell.

11 Problem och fallgropar

Det finns en uppsjö med problem förknippade med regression.

11.1 Stark korrelation

Om två variabler är starkt korrelerade innebär det att matrisen X nästan blir singulär (den blir dåligt **konditionerad**), vilket ställer till det rent numeriskt då avrundningsfel och dylikt nu kan förändra svar drastiskt. Systemet blir helt enkelt väldigt störningskänsligt.
Man brukar undvika starkt korrelerade variabler.
Ett specialfall är när matrisen $X^T X$ inte är inverterbar. Då behöver någon/några variabler tas bort.

11.2 Extrapolation

När vi har våra uppmätta data så får vi direkt ett rätblock i $\mathbf{R}^k$ där

$$x_i^- \leq x_i \leq x_i^+, \quad i = 1, 2, \ldots, k.$$

Talen $x_i^\pm$ är helt enkelt max och min vid mätningen för den uppmätta variabeln x_i. Mellan dessa gränser undersöker vi en linjär regressionsmodell. Denna modell bör inte okvalificerat användas för att uttala sig (prediktera) något utanför rätblocket.

11.3 Residualfördelning

Se till att göra några undersökningar om residualerna. Om de uppvisar ett mönster är det ett tecken på att felen inte uppfyller de krav vi ställt. Om inte felen är normalfördelade (med samma varians) så leder detta till att *samtliga* tester (F-test, varianstest, test för $\beta_i = 0$ etc) inte är tillförlitliga.