F2 Regressionsanalys, introduktion

Vi har en uppsättning variabler

$$Y \quad \text{och} \quad X_1, X_2, \ldots, X_k$$

där Y sägs vara den *beroende* variabeln och $X_1, X_2, \ldots, X_k$ sägs vara de *oberoende*.

Vi vill fastställa hur det *funktionella sambandet* mellan dessa ser ut, dvs

$$Y = f(X_1, X_2, \ldots, X_k)$$

Syftet är att skapa en *modell* som i något avseende beskriver verkligheten på ett bra sätt.

1. Prediktiva modeller, tex Newtons gravitationsteori, Boyles gaslag

2. Förklarande modeller, tex Darwins evolutionsmodell

3. Preskriptiva modeller, tex Keynesianska ekonomiska modeller

Vad är en "bra" modell? Olika kriterier som brukar omtalas är

- Enkelhet (*simplicity*), dvs om en komplicerad modell ger samma svar som en enklare modell, skall den enklare väljas (Occams rakkniv)

- Klarhet, tydlighet (*clarity*), det ska vara lätt att förstå hur modellen ska användas och den ska med "lätthet" producera svar när samma frågor ställs.

- Objektivitet, (*bias free*), min politiska hemvist eller hur jag mådde i morse ska inte påverka resultatet.

- Användbarhet, (*tractability*) om kostnaden för att få svar på frågor med hjälp av modellen är för hög, är det inte en bra modell.

Kausalitet

Vad är det som påverkar vad? I vilken riktning går det?

Ex) Jag behandlas med ett preparat och mitt höga blodtryck sjunker. Var det medicineringen som påverkade blodtrycket eller var det mitt redan sjunkande blodtryck som fick mig att ta medicinen?

Ex) Beror arbetslösheten på inflationen eller inflationen på arbetslöshet?

Ex) Godsfrakt via tåg har visats vara negativt korrelerad med starkölsförsäljningen. Varför?

Viktigt att skilja på *statistiska* (numeriska) samband och *kausala* samband. Statistiska samband är inget bevis för kausala samband! Ibland är starka numeriska samband endast *spuriösa*.

Kriteterier för kausalitet, se sid 37, sju punkter.

Regressionsanalys

Metod för att söka *samband* mellan variabler, och *kvantifiera* detta samband dvs ge ett numeriskt mått på sambandet. Det är ett försök att beskriva verkligheten genom att formulera en *matematisk modell*. Modellen kan sedan, baserat på *observationer*, antingen *accepteras* eller *förkastas*. Dessutom får vi ett verktyg att kvantifiera *osäkerheten* i de slutsatser vi drar (jmfr skillnaden mellan deterministiska och stokastiska modeller, sid 37-38).

Vi skiljer på variablerna

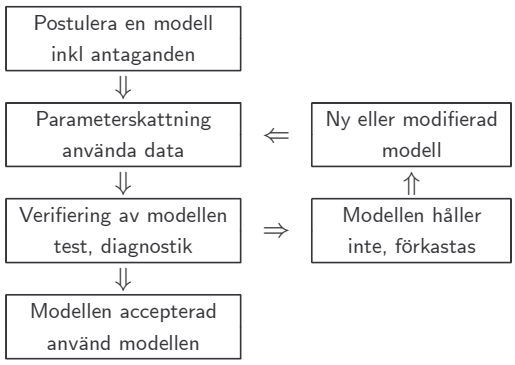
$$Y = \begin{cases} \text{beroende variabel} \\ \text{undersöknings variabel} \\ \text{resultatvariabel} \\ \text{responsvariabel} \end{cases}$$

och

$$X_1, \ldots, X_k = \begin{cases} \text{oberoende variabler} \\ \text{förklaringsvariabler} \\ \text{bakgrundsvariabler} \\ \text{prediktorer} \end{cases}$$

En prediktor $\Longrightarrow$ enkel regression
Flera prediktorer $\Longrightarrow$ multipel regression

Arbetsgången är ofta en *iterativ* process:



Enkel linjär regressionsanalys

Enklaste modellantagandet

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

dvs en *linjär* modell där

- β_0, β_1 är okända parametrar som skattas
- ε är ett slumpfel eller slumpvariabel med en fördelning.
- X är en prediktor för Y. Är ofta betraktad som fix (men kan vara slumpmässig.)
- Y är responsvariabeln som beror på X. Slumpvariabel.

Antaganden:

Existens: För varje fixt värde på X gäller att Y är en slumpvariabel med en fördelningsfunktion med ändligt väntevärde $\mu_{Y|X}$ och ändlig varians $\sigma^2_{Y|X}$.

(Notera notationen!)

Oberoende (independence). Y-värdena i ett stickprov är statistiskt oberoende $\Longleftrightarrow$ slumptermerna ε som associeras med de enskilda Y-värdena, är oberoende. Vidare är ε och X oberoende (om X är en sv).

Linjäritet: Det funktionella sambandet mellan X och Y är linjärt, dvs är på formen

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

eller

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

Homoskedasticitet eller lika varians: För varje fixt värde på X gäller att

$$Var\,(Y \mid X) = \sigma^2_{Y|X} = \sigma^2$$

dvs hur vi än väljer X så är den *betingade* variansen för Y lika stor. Ibland bra att beteckna med σ^2_ε.

OBS! Ej att förväxla med den *obetingade* variansen för Y, dvs σ^2_Y.

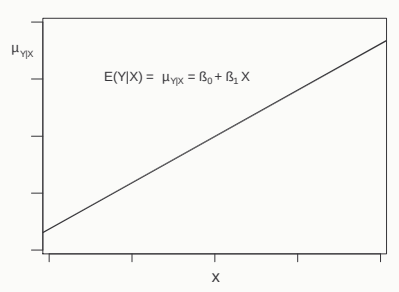
Normalitet: För varje fixt värde på X gäller att Y är normalfördelat med väntevärde $\mu_{Y|X}$ och ändlig varians $\sigma^2_{Y|X}$.

Det betingade väntevärdet för Y kan skrivas

$$\mu_{Y|X} = E\,(Y \mid X = x) = \beta_0 + \beta_1 x$$

och under antagandena är den betingade fördelningen för Y alltså

$$(Y \mid X = x) \sim N\left(\beta_0 + \beta_1 x\ ,\ \sigma^2_{Y|X}\right)$$



Det är alltså de *betingade väntevärdena* för Y som beskrivs av regressionslinjen

$$\mu_{Y|X} = E\left(Y \mid X = x\right) = \beta_0 + \beta_1 x$$

Överallt samma *betingade varians*

$$\sigma^2_{Y|X} = Var\left(Y \mid X = x\right) = \sigma^2_\varepsilon$$

Samt den betingade *fördelningen* för Y givet X, är normal.

Parameterskattning

Vi antar att vi har ett iid stickprov, av storlek n, av observationspar

$$(x_i, y_i)$$

Antag att $\hat{\beta}_0$ och $\hat{\beta}_1$ är två skattningar av regressionskoefficienterna. Vi kan då skriva

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

där $\hat{y}_i$ blir en skattning av $\mu_{Y|X}$, dvs det betingade väntevärdet givet $X = x_i$.

Detta värde $\hat{y}_i$ kan då jämföras med det faktiskt observerade värdet y_i.

Vi kommer ihåg MK-metoden (eng. Least Squares, eller LS).

$$Q = \sum_{i=1}^{n} \left(y_i - \hat{\mu}_Y\right)^2$$

Det värde på $\hat{\mu}_Y$ som minimerar Q är vår skattning för μ_Y.

Istället för att skatta μ_Y skall vi skatta de *betingade* väntevärdena $\mu_{Y|X}$ som ju bestäms av β_0 och β_1.

Skattningarna $\hat{\beta}_0$ och $\hat{\beta}_1$ bestäms så att

$$Q = \sum_{i=1}^{n} \left(y_i - \mu_{Y|X}\right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

minimeras.

Max-min-problem! Derivera Q map $\hat{\beta}_0$ och $\hat{\beta}_1$ och sätt lika med 0, ger

$$\begin{cases} \frac{\partial Q}{\partial \hat{\beta}_0} = -2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\ \frac{\partial Q}{\partial \hat{\beta}_1} = -2\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum y_i - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 x_i = 0 \\ \sum x_i y_i - \sum \hat{\beta}_0 x_i - \sum \hat{\beta}_1 x_i^2 = 0 \end{cases}$$

$$\Rightarrow \begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i & (1) \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i & (2) \end{cases}$$

Från (1) får vi efter lite förenklig

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Notera att paret av medelvärden $(\bar{x}, \bar{y})$ alltid ligger exakt på linjen!

$$\hat{y}_{x=\bar{x}} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

Insättning av detta i (2) ger efter (lite) förenklig

$$\hat{\beta}_1 = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2} = \frac{S_{xy}}{S_x^2}$$

Den sista kan jämföras med formeln för korrelationskoefficienten som kan skrivas

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}\sqrt{\sum\left(y_i - \bar{y}\right)^2}} = \frac{S_{xy}}{S_x \cdot S_y}$$

och vi ser att

$$r = \frac{S_x}{S_y}\hat{\beta}_1 \quad \text{och} \quad \hat{\beta}_1 = \frac{S_y}{S_x} r$$

Kom ihåg att korrelationskoefficienten är ett mått på det *linjära sambandet*!

Formlerna för $\hat{\beta}_1$ och r kan även skrivas som

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - (\sum y_i)(\sum x_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

respektive

$$r = \frac{n\sum x_i y_i - (\sum y_i)(\sum x_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

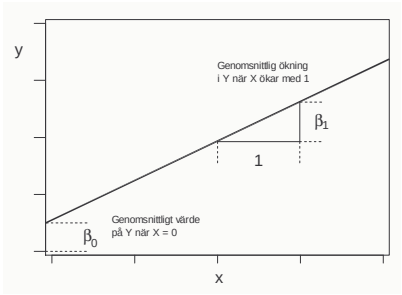
$$\hat{\beta}_1 = \frac{S_Y}{S_X} \cdot r = \frac{S_{XY}}{S_X^2}$$

Vilket man använder är en smakfråga. Idag används i vilket fall datorer!

Tolkning av regressionskoefficienterna:

* $\hat{\beta}_0$ är interceptet, det *genomsnittliga* värdet på y när $x = 0$.

* $\hat{\beta}_1$ är linjens lutning, den *genomsnittliga* förändringen då x ändras en enhet.



Denna förväntade förändring är lika för *alla* värden på x. För ett visst värde x_0 på x har vi

$$\hat{y}_{x=x_0} \;=\; \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Om vi ökar detta värde x_0 med ett får vi

$$\hat{y}_{x=x_0+1} \;=\; \hat{\beta}_0 + \hat{\beta}_1 (x_0 + 1)$$

och ökningen i y-led blir

$$\hat{y}_{x=x_0+1} \;-\; \hat{y}_{x=x_0} \;=\; \hat{\beta}_1$$

Residualer

Från *modellen* har man *slumptermerna*

$$\varepsilon_i \;=\; Y_i - \hat{Y}_i \;=\; Y_i - \mu_{Y|X} \;=\; Y_i - \beta_0 - \beta_1 X_i$$

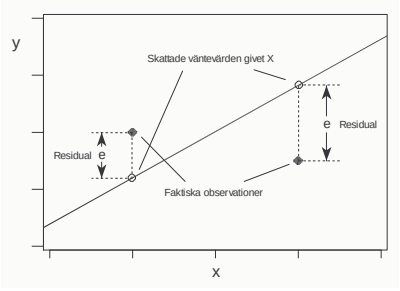
och antagande kan delvis omformuleras enligt

$$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2\right)$$

Från *data* beräknas *residualerna*

$$e_i \;=\; y_i - \hat{y}_i \;=\; y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

för alla i, och residualen e_i ses som en skattning av slumpfelet ε_i.



Enkelt att se att summan av residualer *alltid* är lika med noll:

$$\begin{aligned}\sum_{i=1}^{n} e_i \;&=\; \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \\ &=\; \sum_{i=1}^{n}\left(y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i\right) \\ &=\; \sum_{i=1}^{n} y_i - n\bar{y} + n\hat{\beta}_1\bar{x} - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0\end{aligned}$$

Detta garanteras av MK-metoden. Men då är de inte heller oberoende!

Sum-of-Squared-Errors (SSE)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Om $SSE = 0 \implies$ perfekt anpassning, *alla* residualer lika med noll.

Under givna antaganden har vi en skattning för slumpfelet ε_i i e_i.

En skattning för $\sigma^2_\varepsilon = \sigma^2_{Y|X}$ baseras på SSE.

$$S^2_e = S^2_{Y|X} = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Jämför med den obetingade variansen för Y,

$$S^2_Y = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

som en skattning av σ^2_Y.

Uppdelning i varianskomponenter:

$$\begin{aligned} & (n-1)\, S^2_Y \\ = \; & \sum (y_i - \bar{y})^2 = \sum \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right]^2 \\ = \; & \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \end{aligned}$$

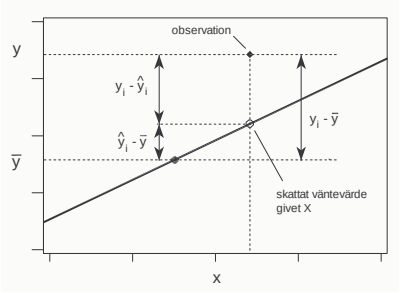
Den sista termen är noll:

$$\begin{aligned} \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum e_i (\hat{y}_i - \bar{y}) \\ &= \sum \left[e_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) - e_i \bar{y} \right] \\ &= \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum e_i x_i - \bar{y} \sum e_i \\ &= \hat{\beta}_1 \sum e_i x_i \end{aligned}$$

Att den sista termen är noll är inte helt lätt att se (se sista sidan).

Vi får

$$\begin{array}{ccccc} \sum (y_i - \bar{y}_i)^2 & = & \sum (y_i - \hat{y}_i)^2 & + & \sum (\hat{y}_i - \bar{y})^2 \\ SST & = & SSE & + & SSR \\ \text{(total variation i } Y\text{)} & & \text{(oförklarad variation)} & & \text{(förklarad variation)} \end{array}$$



Andelen förklarad variation

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

Mål: Få så stor andel förklarad variation som möjligt.

R^2 begränsas enligt

$$0 \leqslant R^2 \leqslant 1 \quad \text{alt} \quad 0 \leqslant R^2 \leqslant 100\%$$

Skattningarna $\hat{\beta}_0$ och $\hat{\beta}_1$ är slumpvariabler. Vad har de för egenskaper med avseende på väntevärde, varians och fördelning?

Kan visas att

$$E\left(\hat{\beta}_1\right) = \beta_1, \quad \text{alltså vvr}$$

$$Var\left(\hat{\beta}_1\right) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\varepsilon}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2_\varepsilon}{(n-1)\, S^2_x}$$

och

$$E\left(\hat{\beta}_0\right) = \beta_0, \quad \text{alltså vvr}$$

$$Var\left(\hat{\beta}_0\right) = \sigma^2_{\hat{\beta}_0} = \sigma^2_\varepsilon \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\, S^2_x} \right)$$

Skattningarna $\hat{\beta}_0$ och $\hat{\beta}_1$ brukar ibland kallas BLUE (*best linear unbiased estimators*)

Vidare har vi under de givna antagandena:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right)$$

och

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$$

Haken är att σ^2_ε är okänd och måste skattas. Men vi hade ju en skattning ovan...

$$S^2_e = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2}$$

Kan visas att $E\left(S^2_e\right) = \sigma^2_\varepsilon$, dvs den är vvr.

Vi sätter alltså in S^2_e istället för σ^2_ε i formlerna för $\sigma^2_{\hat{\beta}_0}$ och $\sigma^2_{\hat{\beta}_1}$ ovan för att få skattningarna

$$S^2_{\hat{\beta}_0} = S^2_e\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)\,S^2_x}\right)$$

och

$$S^2_{\hat{\beta}_1} = \frac{S^2_e}{(n-1)\,S^2_x}$$

Eftersom vi skattar σ^2_ε med S^2_e så används t-fördelningen för inferensen (se sid 54).

Jämför med inferensen när vi skattar μ i fallet med normalfördelning med okänd varians.

Hypotesprövning:

Nollhypotes: $H_0 : \beta_1 = \beta_1^{(0)}$

Testfunktion: $\dfrac{\hat{\beta}_1 - \beta_1^{(0)}}{S_{\hat{\beta}_1}} \sim t\,(n-2)$

Nollhypotes: $H_0 : \beta_0 = \beta_0^{(0)}$

Testfunktion: $\dfrac{\hat{\beta}_0 - \beta_0^{(0)}}{S_{\hat{\beta}_0}} \sim t\,(n-2)$

Konfidensintervall:

$$\hat{\beta}_1 \pm t^{(n-2)}_{1-\alpha/2} \cdot S_{\hat{\beta}_1}$$

och

$$\hat{\beta}_0 \pm t^{(n-2)}_{1-\alpha/2} \cdot S_{\hat{\beta}_0}$$

För den intresserade: bevis för att

$$\sum e_i x_i = 0$$

Använd att $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, och $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ samt att man kan skriva

$$\hat{\beta}_1 = \frac{S_y}{S_x} r = \frac{S_{xy}}{S^2_x}$$

där S_{xy} är kovariansen mellan X och Y i stickprovet. Detta ger

$$\begin{aligned}
\sum e_i x_i &= \sum (y_i - \hat{y}_i)\, x_i = \sum \left(x_i y_i - \left(\hat{\beta}_0 x_i + \hat{\beta}_1 x_i^2\right)\right) \\
&= \sum x_i y_i - \bar{y} \sum x_i + \hat{\beta}_1 \bar{x} \sum x_i - \hat{\beta}_1 \sum x_i^2 \\
&= \left(\sum x_i y_i - n\bar{x}\bar{y}\right) - \hat{\beta}_1 \left(\sum x_i^2 + n\bar{x}^2\right) \\
&= (n-1)\, S_{xy} - \hat{\beta}_1 \cdot (n-1)\, S^2_x \\
&= (n-1)\, S_{xy} - \frac{S_{xy}}{S^2_x}(n-1)\, S^2_x = 0
\end{aligned}$$

F3 Regressionsanalys, forts

Kom ihåg antagandena, 5 stycken:

Existens, Linjäritet, Normalitet, Oberoende, Homoskedasticitet (lika varians)

Att skriva

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \varepsilon_i$$

där $|\beta_0|$ och $|\beta_1| < \infty$, och

$$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma_\varepsilon^2\right), \quad 0 \leqslant \sigma_\varepsilon^2 < \infty$$

sammanfattar antagandena.

Givet ett iid stickprov av observationspar (x_i, y_i) skattas regressionsparametrarna enligt MK-metoden (LS, OLS)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_y}{S_x} r = \frac{S_{xy}}{S_x^2}$$

där S_{xy} är kovariansen i stickprovet.

Vi får *residualer* enligt

$$e_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Den totala variationen i y kan uttryckas som

$$SST = \sum (y_i - \bar{y})^2 = (n-1)\, S_y^2$$

och denna kan delas upp i varianskomponenter

$$SST = SSE + SSR = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

<se bild residualer från F17>

Den del av variationen i Y som *förklaras* av regressionsmodellen är SSR.

Den återstående delen SSE, är den *oförklarade* variationen, den betingade variationen enligt σ_ε^2 och som skattas med

$$S_e^2 = \frac{SSE}{n-2}$$

Andelen förklarad variation skrivs då

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

kallas även *förklaringsgrad*. Notera att

$$0 \leqslant R^2 \leqslant 1$$

Eftersom skattningarna $\hat{\beta}_0$ och $\hat{\beta}_1$ är definierade som linjärkombinationer av normalfördelade slumpvariabler blir även

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma_{\hat{\beta}_0}^2\right) \quad \text{och} \quad \hat{\beta}_1 \sim N\left(\beta_1, \sigma_{\hat{\beta}_1}^2\right)$$

där $\sigma_{\hat{\beta}_0}^2$ och $\sigma_{\hat{\beta}_1}^2$ beror på σ_ε^2.

Men eftersom σ_ε^2 är okänd och måste skattas med S_e^2 används t-fördelningen med $(n-2)$ frihetsgrader för inferensen för β_0 och β_1 (konfidensintervall och hypotesprövning).

Konfidens- och Prediktionsintervall

Antag att $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$. Vi förväntar oss att se ca $(1-\alpha)\,100\%$ av alla observationer i intervallet

$$\mu_Y \ \pm \ z_{1-\alpha/2} \cdot \sigma_Y^2$$

Med okända men skattade parametervärden

$$\hat{\mu}_Y = \bar{y} \quad \text{och} \quad \hat{\sigma}_Y^2 = S_y^2$$

blir vår bästa gissning om framtida / icke-observerade värden på Y, ett $(1-\alpha)\,100\%$ sk *prediktionsintervall* (PI), given enligt

$$\bar{y} \ \pm \ t_{1-\alpha/2}^{(n-1)} \cdot S_y$$

Variansen för Y är alltid σ_Y^2, oavsett hur många y vi ser.

Ett $(1-\alpha)\,100\%$ *konfidensintervall* (KI) däremot säger något om hur bra skattningen av μ_Y är. Precisionen i skattningen blir bättre ju större stickprovsstorlek

$$\bar{y} \ \pm \ t_{1-\alpha/2}^{(n-1)} \cdot \frac{S_y}{\sqrt{n}}$$

Variansen för stickprovsmedelvärdet, $Var\left(\bar{Y}\right) = \sigma_Y^2/n$, blir mindre ju större stickprov vi har, vi har bättre precision i skattningen av μ_Y.

Motsvarigheten för en skattad enkel linjär regressionsmodell blir:

Inferens om betingat väntevärde $\mu_{Y|X}$ givet att $X = x_0$ ges av skattningen

$$\hat{\mu}_{Y|X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

och *konfidensintervallet*

$$\hat{\mu}_{Y|X=x_0} \ \pm \ t_{1-\alpha/2}^{(n-2)} \cdot S_e \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\,S_x^2}}$$

Notera att om vi beräknar detta för $x_0 = \bar{x}$ får vi

$$\bar{y} \ \pm \ t_{1-\alpha/2}^{(n-2)} \cdot \frac{S_e}{\sqrt{n}}$$

Prediktion av "framtida" Y givet $X = x_0$ ges av skattningen

$$\hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

och *prediktionsintervallet*

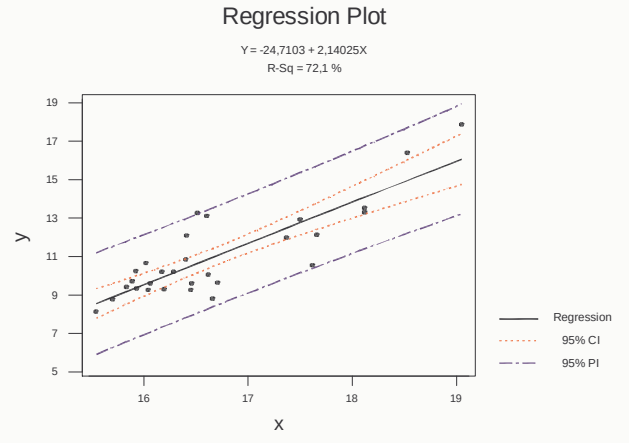
$$\hat{Y}_{x_0} \ \pm \ t_{1-\alpha/2}^{(n-2)} \cdot S_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\,S_x^2}}$$

Notera att om vi beräknar detta för $x_0 = \bar{x}$ får vi

$$\bar{y} \ \pm \ t_{1-\alpha/2}^{(n-2)} \cdot S_e \cdot \sqrt{1 + \frac{1}{n}}$$

$$\approx \ \bar{y} \ \pm \ t_{1-\alpha/2}^{(n-2)} \cdot S_e$$

Med Minitab



Kap 6 Korrelationskoefficienten

Korrelationskoefficienten r för stickprovet definieras

$$\begin{aligned} r &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}} = \frac{S_{xy}}{S_x S_y} = \frac{S_x}{S_y}\hat{\beta}_1 \end{aligned}$$

Egenskaper:

- $-1 \leqslant r \leqslant 1$
- skaloberoende
- r och $\hat{\beta}_1$ har samma tecken
- r ger ett mått på styrkan i det *linjära* sambandet

Allmänt gäller att

$$r^2 = R^2 = \frac{SSY - SSE}{SSY}$$

Notera också att man ofta ser olika beteckningar för samma sak:

$$SSy = SSY = SST = \sum (y_i - \bar{y})^2$$

För en bivariat normalfördelning

$$(X, Y) \sim BVN\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho\right)$$

kan regressionsmodellen skrivas som

$$\mu_{Y|X} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_x)$$

vilket kan jämföras med motsvarande skattning av $\mu_{Y|X}$ med den skattade modellen

$$\begin{aligned} \hat{\mu}_{Y|X} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1(x - \bar{x}) \\ &= \bar{y} + r\frac{S_Y}{S_X}(x - \bar{x}) \end{aligned}$$

Inferens om ρ

1. $H_0 : \rho = 0 \quad \Leftrightarrow \quad H_0 : \beta_1 = 0$ med teststatistika

$$T = \frac{r \cdot \sqrt{n-2}}{\sqrt{1 - r^2}} \sim t(n-2)$$

2. $H_0 : \rho = \rho_0$ där $\rho_0 \neq 0$

- ej ekvivalent med $H_0 : \beta_1 = \beta_1^{(0)}$ för något $\beta_1^{(0)} \neq 0$
- teststatistikans fördelning symmetrisk endast om $\rho_0 = 0$ som i 1 ovan. Då $\rho_0 \neq 0$ är fördelningen skev.
- exakt fördelning för r är komplicerad.

$\Longrightarrow$ Transformera till något approx. normalfördelat!

Fishers Z-transformation

Vi definierar

$$\hat{z} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad \text{och} \quad z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

och får en teststatistika

$$Z = (\hat{z} - z_0) \cdot \sqrt{n-3} \qquad (6.9)$$

som är approximativt $N(0,1)$ under H_0.

Konfidensintervall för ρ

Använd Fishers Z-transformation enligt

$$L_Z = \frac{1}{2}\ln\left(\frac{1+L_\rho}{1-L_\rho}\right) \quad \text{och} \quad U_Z = \frac{1}{2}\ln\left(\frac{1+U_\rho}{1-U_\rho}\right)$$

där L_Z och U_Z är nedre resp övre gräns i intervallet

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \pm \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \qquad (6.10)$$

och där $z_{1-\alpha/2}$ är den vanliga normalfördelningens kvantil, tex 1.96 (dubbelsidigt KI)

Lös sedan ut L_ρ och U_ρ enligt

$$L_\rho = \frac{e^{2L_Z}-1}{e^{2L_Z}+1} \quad \text{och} \quad U_\rho = \frac{e^{2U_Z}-1}{e^{2U_Z}+1}$$

Alltså, beräkna först (6.10) och sedan ovanstående för att få ett intervall för ρ.

Om detta intervall sedan täcker in ρ_0 enligt nollhypotesen, accepteras H_0.

3. Man kan jämföra korrelationskoefficienterna för två oberoende stickprov för att se om de är lika dvs

$$H_0 : \rho_1 = \rho_2$$

Detta kan ni läsa kursivt.

Mer inferens i regressionsanalys

Vi är intresserade av att förklara variationen i Y med X. Andelen förklarad variation ges ju av

$$R^2 = 1 - \frac{SSE}{SST}$$

$R^2 = 1$ (100%)

$\iff \quad SSE = 0$ dvs $\sum (y_i - \hat{y}_i)^2 = 0$ och samtliga observationer ligger exakt på linjen.

$R^2 = 0$

$\iff \quad SSR = 0$ dvs $\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2$ regressionslinjen ligger parallellt med x-axeln. Vi har $\hat{y}_i = \bar{y}$ för alla i.

Vi vill testa om SSR är "stor" i förhållande till SSE, dvs om det "lönar sig" att anpassa en modell.

Vad har SSR och SSE för fördelning? Definiera

$$MSR = SSR/1 \quad (1 \text{ prediktor})$$

$$MSE = SSE/(n-2) = S_e^2$$

Det kan visas att

$$E(MSR) = \sigma_\varepsilon^2 + \underbrace{\beta_1^2 \cdot \sum (x_i - \bar{x})^2}_{>0}$$

$$E(MSE) = \sigma_\varepsilon^2$$

- Om $\beta_1 = 0$, dvs regression inte föreligger, så är $E(MSR) = E(MSE) = \sigma_\varepsilon^2$
- Om $\beta_1 \neq 0$ så är $E(MSR) > E(MSE)$

Fördelning under en nollhypotes $H_0 : \beta_1 = 0$:

$$\frac{SSR}{\sigma_\varepsilon^2} \sim \chi^2(1)$$

och

$$\frac{SSE}{\sigma_\varepsilon^2} \sim \chi^2(n-2)$$

Det kan dessutom visas att dessa är oberoende. Vi får

$$\frac{MSR}{MSE} \sim F(1, n-2) \quad \text{då} \quad \beta_1 = 0$$

Vi kan testa om modellens förklaringsgrad är tillräckligt stor, dvs om det *föreligger regression*.

Tre ekvivalenta test

1. Baserat på MSR och MSE :

   H_0 : $\beta_1 = 0$ "ingen regression"
   H_1 : $\beta_1 \neq 0$ "regression"

   Testfunktion

   $$F_{obs} = \frac{MSR}{MSE} \sim F(1, n-2)$$

   Förkasta H_0 om $F_{obs} >$ kritiskt värde, enkelsidigt test.

2. Baserat på skattningen $\hat{\beta}_1$:

   $$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_1 : \beta_1 \neq 0$$

   Testfunktion

   $$T_{obs} = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} \sim t(n-2)$$

   Förkasta H_0 om $|T_{obs}| >$ kritisktvärde, dubbelsidigt test.

3. Baserat på korrelationskoefficienten r :

   $$H_0 : \rho = 0 \quad \text{mot} \quad H_1 : \rho \neq 0$$

   Testfunktion

   $$T_{obs} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$

   Förkasta H_0 om $|T_{obs}| >$ kritisktvärde, dubbelsidigt test.

Det kan vara värt att notera att de kritiska värdena för F- och t-fördelningarna har följande egenskap

$$F_{1-\alpha}^{(1,r)} = \left(t_{1-\alpha/2}^{(r)}\right)^2$$

Ex) För $r = 5$ och $\alpha = 5\%$ får vi

$$F_{0.95}^{(1,5)} = 6.61$$
$$\left(t_{0.975}^{(5)}\right)^2 = (2.571)^2 = 6.61$$

Kap 7. ANOVA-tablån

ANOVA = analysis of variance, dvs analys av varianskomponenterna SST, SSR och SSE.

Källa	Kvadratsumma	Frihetsgrader	Medelkvadratsumma
Regression	SSR	1	$MSR = \frac{SSR}{1}$
Residual (error)	SSE	$(n-2)$	$MSE = \frac{SSE}{(n-2)} = S_e^2$
Total	SST	$(n-1)$	

Vi kommer ihåg att

1 : $SST = \sum (y_i - \bar{y})^2$
2 : $SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$
3 : $SSR = \sum (\hat{y}_i - \bar{y})^2$

där

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

och

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Vi har att

1 : $SST = SSR + SSE$

2 : $\dfrac{SST}{(n-1)} = MST = S_y^2$

3 : $MST \neq MSR + MSE$

3 : $(n-1) = 1 + (n-2)$

4 : $F_{obs} = \dfrac{MSR}{MSE}$, ($F$-kvoten för test)

Jämför ANOVA-tablån ovan med Minitab-utskriften vid en regressionsanalys.

.

.

.

F4 Regressionsanalys, forts

Kap 8. Multipel regression

Utöka den enkla linjära regressionsmodellen till en linjär modell med fler prediktorer:

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}_{k \text{ st prediktorer}}$$

Saker att ta hänsyn till:

- val av modell kan vara svårare (vilka prediktorer ska man ha med?)
- svårt att grafiskt åskådliggöra modellen (plotta data i fler än 2-3 dimensioner)
- tolkning av regressionkoefficienterna
- beräkningarna blir komplicerade, kräver ofta dator

Antaganden

Existens: För varje fixt värde på resp $X_1, \ldots, X_k$ gäller att Y är en slumpvariabel med ändligt väntevärde $\mu_{Y|X_1 \ldots X_k}$ och ändlig positiv varians $\sigma^2_{Y|X_1 \ldots X_k}$.

Oberoende (independence). Y-värdena i ett stickprov är statistiskt oberoende $\Longleftrightarrow$ slumptermerna ε som associeras med de enskilda Y-värdena, är oberoende. Vidare är ε och $X_1, \ldots, X_k$ oberoende (om $X_1, \ldots, X_k$ är sv).

Linjäritet: Det funktionella sambandet mellan $X_1, .., X_k$ och Y är linjärt, dvs

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

och

$$\mu_{Y|X_1 \ldots X_k} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Homoskedasticitet eller lika varians: För varje fixt värde på resp $X_1, \ldots, X_k$ gäller att

$$\sigma^2_{Y|X_1 \ldots X_k} = \sigma^2_\varepsilon$$

dvs hur vi än väljer värden på resp $X_1, \ldots, X_k$ så är den betingade variansen för Y lika stor.

Normalitet: För varje fixt värde på resp $X_1, \ldots, X_k$ gäller att Y är normalfördelat med väntevärde $\mu_{Y|X_1 \ldots X_k}$ och ändlig varians $\sigma^2_{Y|X_1 \ldots X_k}$.

Vi sammanfattar detta som att

$$Y_i \mid X_{1i}, \ldots, X_{ki} \;=\; \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

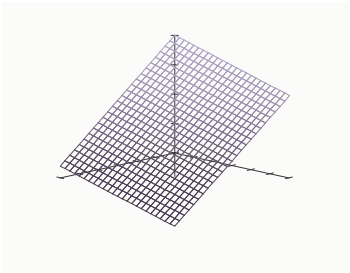
där $|\beta_i| < \infty$ för $i = 0, 1, \ldots, k$ och att

$$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2_\varepsilon\right), \quad 0 \leqslant \sigma^2_\varepsilon < \infty$$

Ex) Multipel modell med två prediktorer, tre dimensioner

$$Y = 10 - 3X_1 - X_2$$

ger en *regressionsyta* eller ett *regressionsplan*:



Det kommer att finnas *observerade* punkter (x_{1i}, x_{2i}, y_i) ovan- och nedanför detta plan.

Vi vill hitta det regressionsplan som minimerar summan av de kvadrerade avstånden mellan observationer och planet, dvs

$$\sum (y_i - \hat{y}_i)^2 = \sum \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \right) \right]^2 = \sum e_i^2$$

dvs Minsta-Kvadratskattningen av $\hat{\beta}_0, \hat{\beta}_1$ och $\hat{\beta}_2$.

Inte längre några lätta formler för skattningarna. För den intresserade: skattningarna kan enkelt uttryckas på matrisform enligt

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y}$$

men detta kräver ytterligare matematikkunskaper, se tex Sydsäter kap 12.

Egenskaper:

- Estimatorerna $\hat{\beta}_j$ är linjärkombinationer av normalfördelade sv (Y-värdena) $\Longrightarrow$

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right), \quad \text{för } j = 0, 1, \ldots, k.$$

- Den MK-skattade regressionsmodellen $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$ är den linjärkombination av X_j'na som ger den starkaste korrelationen (närmast ± 1) med Y. Dvs $Corr\left(Y, \hat{Y}\right)$ är det starkaste vi kan hitta för olika val på $\beta_0, \beta_1, \ldots, \beta_k$.

- Enkel linjär regression motsvaras av en bivariat normalfördelning. Multipel linjär regression motsvaras av en *multivariat* normalfördelning.

ANOVA-tablån för multipel regressionmodell (jämför med Minitab-utskriften vid en multipel regressionsanalys):

Källa	Kvadrat-summa	Frihets-grader	Medelkvad-ratsumma
Regression	SSR	k	$MSR = \frac{SSR}{k}$
Residual (error)	SSE	$(n-k-1)$	$MSE = \frac{SSE}{(n-k-1)} = S_e^2$
Total	SST	$(n-1)$	

Det som sades tidigare om enkel linjär regression gäller fortfarande, tex att $SST = SSR + SSE$.

Kom ihåg också att

$$MSE = \frac{SSE}{n-k-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-k-1} = S_e^2 = \hat{\sigma}_\varepsilon^2$$

Förklaringsgrad, andel förklarad variation definieras fortfarande som

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Problemet med R^2 är att den *alltid* ökar när man utökar modellen med fler prediktorer och därmed inte är ett bra kriterium för jämförelser mellan "mindre" modeller (få prediktorer) och "större" modeller (många prediktorer).

Genom att ta hänsyn till att vi har färre frihetsgrader med en "större modell" kan man istället studera det *justerade* R^2-värdet som definieras enligt

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{MSE}{MST} = 1 - \frac{S_e^2}{S_y^2}$$

Detta värde kan ibland faktiskt *minska* när man tillför en prediktor.

Tolkning av regressionskoefficienterna:

1. $\hat{\beta}_0$ är det skattade genomsnittliga värdet av Y då alla $X_1 = \ldots = X_k = 0$.

2. $\hat{\beta}_j$ är den skattade genomsnittliga ökningen i Y då X_j ökar med ett, då *alla* övriga X_i, $i \neq j$ hålls *konstanta*.

Ex) Låt $Y = 10 - 3X_1 - X_2 + \varepsilon$. Då är

$$Y = 10 + \varepsilon \quad \text{då} \quad X_1 = X_2 = 0$$

Håll $X_2 = x_2$ konstant. Då är

$$\hat{Y}_{x_1} = 10 - 3X_1 - x_2$$

och

$$\hat{Y}_{x_1+1} = 10 - 3(X_1 + 1) - x_2$$

och

$$\hat{Y}_{x_1+1} - \hat{Y}_{x_1} = -3(X_1 + 1) + 3X_1 = -3$$

och analogt om vi håller $X_1 = x_1$ konstant.

Kap 9. Multipelregression och inferens

Vi kommer huvudsakligen att behandla tre olika varianter av hypotestest:

1. Allmänt test, test på *hela* modellen (overall). Detta test tittar *inte* på enskilda prediktorer och deras koefficienter utan försöker avgöra om modellen som helhet ger en signifikant ökning av förklaringsgraden jämfört med ingen modell alls.

   H_0 : ingen regression (modellen håller ej) mot
   H_1 : regression (modellen håller)

   vilket är ekvivalent med

   H_0 : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ mot
   H_1 : *minst* en av $\beta_1, \beta_2, \ldots, \beta_k$ är $\neq 0$

   Testvariabel:

   $$F = \frac{MSR}{MSE} \sim F(k, n-k-1)$$

   Förkasta om $F_{obs} > F_{1-\alpha}^{(k,n-k-1)}$.

2. Partiellt F-test. Vi har en modell med säg p stycken prediktorer, $X_1, .., X_p$. Nu finns ytterligare en kandidat, säg X^*, som vi funderar på att utöka modellen med. Frågan är om denna kommer att ge en signifikant ökning i förklaringsgrad, givet att $X_1, \ldots, X_p$ redan förklarar "sin del".

   H_0 : X^* tillför inget givet de övriga mot
   H_1 : X^* tillför något givet de övriga

   Sättet att testa detta går ut på att se om ökningen i SSR är signifikant eller inte.

Då SSR alltid ökar (jämför med R^2), jämförs den observerade *ökningen* i SSR mot MSE för den utökade modellen.

Vi definierar testvariabeln enligt

$$F(X^* \mid X_1, \ldots, X_p)$$

$$= \frac{SSR(X^* \mid X_1, \ldots, X_p)}{MSE(X_1, \ldots, X_p, X^*)} \sim F(1, n-p-2)$$

där $MSE(X_1, \ldots, X_p, X^*) = MSE$ för *hela* modellen, inklusive X^* som prediktor, och där

$$\underbrace{SSR(X^* \mid X_1, .., X_p)}_{\text{Ökning i } SSR}$$

$$= \underbrace{\overbrace{SSR(X_1, .., X_p, X^*)}^{\text{Utökad modell}} - \overbrace{SSR(X_1, .., X_p)}^{\text{Tidigare modell}}}_{\text{Skillnaden}}$$

$$= \underbrace{\overbrace{SSE(X_1, .., X_p)}^{\text{Tidigare modell}} - \overbrace{SSE(X_1, .., X_p, X_1^*)}^{\text{Utökad modell}}}_{\text{Skillnaden}}$$

Var hittar vi dessa storheter i Minitab-utskriften?

Alternativt och ekvivalent t-test:

Man testar om koefficienten β^* för den nya prediktorn X^* är signifikant skiljt från noll, dvs

$$H_0 : \beta^* = 0 \quad \text{mot} \quad H_1 : \beta^* \neq 0$$

med testvariablen

$$T = \hat{\beta}^* / S_{\hat{\beta}^*}$$

och förkastar om $|t_{obs}| \geqslant t_{1-\alpha/2}^{(n-p-2)}$ som vanligt.

Här utgår man ifrån den "fulla" modellen med samtliga $p+1$ prediktorerna (dvs med $p+2$ parametrar).

Kom ihåg att

$$F_{obs}^{(1,r)} = \left(t_{obs}^{(r)}\right)^2$$

Värdena anges som standard i programutskriften i så gott som samtliga programvaror, jämför med Minitab-utskriften.

3. Multipel partiellt F-test. Detta är en generalisering av testet under 2. Vi kan testa om inte bara en utan om k stycken nya prediktorer tillsammans tillför något till förklaringsgraden. Vi använder

$$\underbrace{SSR\left(X_1^*,..,X_k^* \mid X_1,..,X_p\right)}_{\text{Ökning i } SSR}$$

$$= \underbrace{\overbrace{SSR\left(X_1,..,X_p,X_1^*,..,X_k^*\right)}^{\text{Utökad modell}} - \overbrace{SSR\left(X_1,..,X_p\right)}^{\text{Tidigare modell}}}_{\text{Skillnaden}}$$

$$= \underbrace{\overbrace{SSE\left(X_1,..,X_p\right)}^{\text{Tidigare modell}} - \overbrace{SSE\left(X_1,..,X_p,X_1^*,..,X_k^*\right)}^{\text{Utökad modell}}}_{\text{Skillnaden}}$$

Nollhypotes och mothypotes blir

H_0 : $X_1^*,..,X_k^*$ tillför inget mot
H_1 : minst en av $X_1^*,..,X_k^*$ tillför något

Testvariabel blir

$$F\left(X_1^*,..,X_k^* \mid X_1,\ldots,X_p\right)$$

$$= \frac{SSR\left(X_1^*,..,X_k^* \mid X_1,\ldots,X_p\right)/k}{MSE\left(X_1,\ldots,X_p,X_1^*,..,X_k^*\right)}$$

$$\sim F\left(k, n-p-k-1\right)$$

k frihetsgrader i täljaren för att vi tillför k nya prediktorer, $n-p-k-1$ frihetsgrader i nämnaren för att vi jämför mot en "full" modell med samtliga $p+k+1$ skattade parametrar.

Alternativt kan vi skriva

$$F = \frac{\left[SSR\left(\text{full}\right) - SSR\left(\text{reducerad}\right)\right]/k}{SSE\left(\text{full}\right)/\left(n-p-k-1\right)}$$

Då $k=1$ har vi testet enligt 2. ovan.

Kan man hitta dessa storheter i Minitab-utskriften?

4. Test för intercept (sid 150):

Inte så vanligt men kan förekomma. Vi testar

$$H_0 : \beta_0 = 0 \quad \text{mot} \quad H_1 : \beta_0 \neq 0$$

Testvariabel blir, under förutsättning att vi har p stycken prediktorer,

$$F = \frac{SSE\left(\text{utan intercept}\right) - SSE\left(\text{med intercept}\right)}{SSE\left(\text{med intercept}\right)}$$
$$\sim F\left(1, n-p-1\right)$$

och vi förkastar för stora värden på F_{obs}

9-5. Strategier för partiella F-test

och hur de kan presenteras i programutskrifter.

Sekvensiellt (ex Minitab, SAS):

Ordningen för hur man anger prediktorerna är avgörande.

Börjar med den först angivna (X_1), sedan den andra (X_2) givet den första (X_1) , tredje (X_3) givet de två första (X_1, X_2) osv.

Ex) ANOVA: med tre prediktorer

Källa		Kvadrat-summa	Frihets-grader
Regr	(X_1)	$SSR(X_1)$	1
	$(X_2 \mid X_1)$	$SSR(X_2 \mid X_1)$	1
	$(X_3 \mid X_1, X_2)$	$SSR(X_3 \mid X_1, X_2)$	1
Residual (error)		SSE	$n-3-1$
Total		SST	$(n-1)$

Sist-in-i-modellen (SAS):

Bedöm varje prediktor som om den var den sista att tillföras modellen, givet de övriga, dvs

Källa		Kvadrat-summa	Frihets-grader
Regr	$(X_1 \mid X_2, X_3)$	$SSR(X_1 \mid X_2, X_3)$	1
	$(X_2 \mid X_1, X_3)$	$SSR(X_2 \mid X_1, X_3)$	1
	$(X_3 \mid X_1, X_2)$	$SSR(X_3 \mid X_1, X_2)$	1

Varje sådant test som vi kan göra på de enskilda prediktorerna och deras respektive koefficienter har då ett alternativt och ekvivalent t-test enligt tidigare beskrivning.

Imorgon går vi igenom exempel på ovanstående test, kom ihåg att ta med Minitab-utskriften!

Tills dess, försök att identifiera de olika kvadratsummorna själva! Vissa behöver räknas ut men allt som behövs finns i utskriften.

F5 Regressionsanalys

Kap 9. Exempel på F- och t-test:

1. Allmänt test på *hela* modellen. Vi har ett stickprov, $n = 50$, och ANOVA-tablån:

Source	DF	SS	MS	F
Regr	3	5499.6	1833.2	99.61
Error	46	846.6	18.4	
Total	49	6346.2		

Hypotes och mothypotes

$$H_0 \ : \ \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{mot}$$
$$H_1 \ : \ \text{minst en av } \beta_1, \beta_2, \beta_3 \text{ är } \neq 0$$

Testvariabel

$$F = \frac{MSR}{MSE} \sim F(k, n-k-1) = F(3, 46)$$

under H_0 och vi förkastar H_0 om vi observerar

$$F_{obs} > F_{1-\alpha}^{(3,46)} \qquad (2.86627 \text{ för } \alpha = 0.05)$$

Vi observerar

$$F_{obs} = \frac{1833.2}{18.4} = 99.61$$

vilket har ett p-värde lika med 0 (enligt Minitab). H_0 förkastas, minst en av $\beta_1, \beta_2, \beta_3$ är $\neq 0$.

2. Partiellt F-test: Fem olika test är möjliga med den givna utskriften; vi kan testa

(a) om $X3$ tillför något givet att $X4$ och $X1$ redan är med

(b) om $X1$ tillför något givet att $X4$ och $X3$ redan är med

(c) om $X4$ tillför något givet att $X1$ och $X3$ redan är med, och

(d) om $X1$ tillför något givet att endast $X4$ redan är med.

(e) om endast $X4$ tillför något till en "tom" modell, $Y = \beta_0 + \varepsilon$.

Vi har från Minitab-utskriften

Predictor	Coef	StDev	T	DF
$X4$	0.40230	0.04418	9.11	46
$X1$	−0.4312	0.4319	−1.00	46
$X3$	2.5518	0.4345	5.87	46

samt den sekvensiella ANOVA-tablån

Source	DF	Seq SS	Boken
$X4$	1	4839.8	$SSR(X4)$
$X1$	1	25.0	$SSR(X1 \mid X4)$
$X3$	1	634.8	$SSR(X3 \mid X4, X1)$
Σ	3	5499.6	$SSR(X4, X1, X3)$

(a) Test om $X3$ tillför något givet att $X4$ och $X1$ redan är med. Testvariabel

$$F(X3 \mid X4, X1) = \frac{SSR(X3 \mid X4, X1)}{MSE(X4, X1, X3)} \sim F(1, n-p-2) = F(1, 46)$$

under H_0 och vi förkastar H_0 om vi observerar

$$F(X3 \mid X4, X1) \geqslant F_{1-\alpha}^{(1,46)}$$

Vi observerar

$$F_{obs} = \frac{634.8}{18.4} = 34.5$$

vilket har ett p-värde lika med 0 (enligt Minitab). H_0 förkastas, $X3$ tillför något

Alternativt t-test med testvariabel

$$T = \frac{\hat{\beta}_3}{S_{\hat{\beta}_3}} \sim t(46)$$

under H_0 och vi förkastar H_0 om vi observerar

$$|t_{obs}| \geqslant t_{1-\alpha/2}^{(46)}$$

(notera absolutbelopp). Vi observerar

$$|t_{obs}| = \left|\frac{2.5518}{0.4345}\right| = 5.87$$

vilket har ett p-värde lika med 0 (enligt Minitab). H_0 förkastas, $X3$ tillför något. Observera också att

$$t_{obs}^2 = 5.87^2 = 34.5 = F_{obs}$$

(b) Test om $X1$ tillför något givet att $X4$ och $X3$ redan är med. Samma som t-testet enligt ovan, vi har

$$T = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t\,(46)$$

och vi har

$$|t_{obs}| = \left|\frac{-0.4312}{0.4319}\right| = 1.00$$

med ett p-värde 0.323, alltså kan H_0 inte förkastas på någon rimlig risknivå.

(c) Test om $X4$ tillför något givet att $X1$ och $X3$ redan är med. Samma som t-testet enligt ovan, vi har

$$T = \frac{\hat{\beta}_4}{S_{\hat{\beta}_4}} \sim t\,(46)$$

och vi har

$$|t_{obs}| = \left|\frac{0.40230}{0.04418}\right| = 9.11$$

vilket har ett p-värde lika med 0 (enligt Minitab). H_0 förkastas, $X4$ tillför något.

(d) Test om $X1$ tillför något givet att endast $X4$ redan är med. Partiellt F-test. Vi måste räkna lite själva först. Testvariabel är

$$\begin{aligned} & F\,(X1 \mid X4) \\ = \; & \frac{SSR\,(X1 \mid X4)}{MSE\,(X4, X1)} = \frac{SSR\,(X1 \mid X4)}{SSE\,(X4, X1)\ /\ 47} \\ = \; & \frac{SSR\,(X1 \mid X4)}{\left[\begin{array}{c} SSE\,(X4, X1, X3) \\ +SSR\,(X3 \mid X4, X1) \end{array}\right]\ /\ 47} \\ \sim \; & F\,(1, 47) \end{aligned}$$

under H_0. Vi förkastar H_0 om vi observerar

$$F\,(X1 \mid X4) \geqslant F_{1-\alpha}^{(1,47)} \qquad (4.047 \text{ för } \alpha = 0.05)$$

Vi observerar

$$\begin{aligned} F_{obs} &= \frac{25.0}{(846.6 + 634.8)\,/47} \\ &= \frac{25.0}{1481.4} = 0.793 \end{aligned}$$

vilket har ett p-värde lika med 0.378 (enligt Minitab). H_0 kan ej förkastas, $X1$ tillför inte något givet att vi redan tagit med $X4$.

(e) Test om endast $X4$ tillför något till en "tom" modell. Vi måste räkna lite själva igen. Testvariabel är

$$\begin{aligned} & F\,(X4 \mid \text{tom}) \\ = \; & \frac{SSR\,(X4 \mid \text{tom})}{MSE\,(X4)} = \frac{SSR\,(X4 \mid \text{tom})}{SSE\,(X4)\ /\ 48} \\ = \; & \frac{SSR\,(X4 \mid \text{tom})}{\left[\begin{array}{l} SSE\,(X4, X1, X3) \\ +\ SSR\,(X3 \mid X4, X1) \\ +\ SSR\,(X1 \mid X4) \end{array}\right]\ /\ 48} \\ \sim \; & F\,(1, 48) \end{aligned}$$

Vi förkastar om vi observerar

$$F\,(X4 \mid \text{tom}) \geqslant F_{1-\alpha}^{(1,48)} \qquad (4.043 \text{ för } \alpha = 0.05)$$

och vi observerar

$$F_{obs} = \frac{4839.8}{(846.6 + 634.8 + 25.0)\ /\ 48} = 154.22$$

3. Multipel partiellt F-test: vi testar om $X1$ och $X3$ tillsammans tillför något givet att $X4$ redan är med i modellen. Testvariabel är

$$\begin{aligned} & F\,(X1, X3 \mid X4) \\ = \; & \frac{SSR\,(X1, X3 \mid X4)\ /\ 2}{MSE\,(X4, X1, X3)} \\ = \; & \frac{SSR\,(X1 \mid X4) + SSR\,(X3 \mid X4, X1)}{MSE\,(X4, X1, X3)} \\ \sim \; & F\,(2, 46) \end{aligned}$$

under H_0 och vi förkastar H_0 om vi observerar

$$F\,(X1, X3 \mid X4) \geqslant F_{1-\alpha}^{(2,46)} \qquad (3.200 \text{ för } \alpha = 0.05)$$

Vi har observerat

$$F_{obs} = \frac{(25.0 + 634.8)\ /\ 2}{18.4} = 17.93$$

vilket har ett p-värde lika med 0 (enligt Minitab). H_0 förkastas, $X1, X3$ tillför något, givet att $X4$ redan finns med i modellen. Men vi vet inte om *båda* tillför något...

Kap 11 Interaktion och samspelseffekter

Ex)

Medicin A har en blodtryckssänkande effekt, α.

Medicin B har en blodtryckssänkande effekt, β.

Var och en för sig kan vi skriva

$$\text{blodtryck efter} = \text{blodtryck innan} + \alpha$$

resp

$$\text{blodtryck efter} = \text{blodtryck innan} + \beta$$

Men kombinerar man A och B är det inte säkert att den totala effekten är additiv, dvs vi kanske istället har en sammantagen effekt enligt

$$\text{blodtryck efter} = \text{blodtryck innan} + \alpha + \beta + \gamma$$

där γ är en *extra* effekt av kombinationen, en samspelseffekt. α och β kallas då för *huvudeffekter* (eng. main effects).

Modellering av samspel i regressionsmodeller:

Modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

1. Antag nu att $X_2 = c$. Då kan modellen skrivas

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + c\beta_2 + \varepsilon \\ &= (\beta_0 + c\beta_2) + \beta_1 X_1 + \varepsilon \\ &= \beta_0^* + \beta_1 X_1 + \varepsilon \end{aligned}$$

2. Nu ökar vi X_2 med ett, dvs $X_2 = c + 1$, och får

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + (c+1)\,\beta_2 + \varepsilon \\ &= (\beta_0 + c\beta_2 + \beta_2) + \beta_1 X_1 + \varepsilon \\ &= (\beta_0^* + \beta_2) + \beta_1 X_1 + \varepsilon \end{aligned}$$

Ingen samspelseffekt, effekten av att ändra X_2 är rent additiv.

För varje val av X_2 så är *lutningen* för X_1 densamma, endast *interceptet* ändras.

Ny modell: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$ med en *multiplikativ* samspelsterm.

1. Antag igen att $X_2 = c$. Då kan modellen skrivas

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + c\beta_2 + c\beta_{12} X_1 + \varepsilon \\ &= (\beta_0 + c\beta_2) + (\beta_1 + c\beta_{12})\, X_1 + \varepsilon \\ &= \beta_0^* + \beta_1^* X_1 + \varepsilon \end{aligned}$$

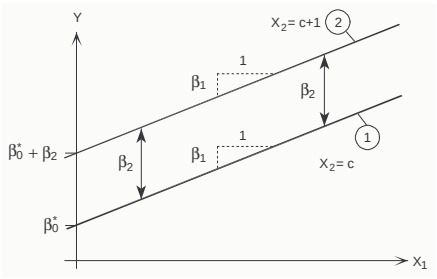
2. Nu ökar vi X_2 med ett, dvs $X_2 = c + 1$, och får

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + (c+1)\,\beta_2 + (c+1)\,\beta_{12} X_1 + \varepsilon \\ &= (\beta_0 + c\beta_2 + \beta_2) + (\beta_1 + c\beta_{12} + \beta_{12})\, X_1 + \varepsilon \\ &= (\beta_0^* + \beta_2) + (\beta_1^* + \beta_{12})\, X_1 + \varepsilon \end{aligned}$$

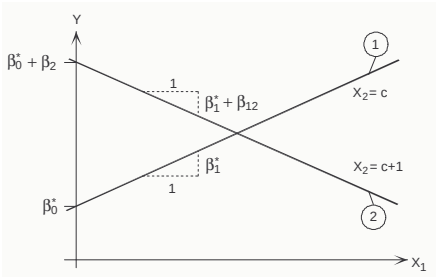
dvs *både* lutning och intercept förändras när vi ändrar på X_2.

Det finns en effekt av att kombinera förändringar i både X_1 och X_2 som inte är rent additiv.

Additiv effekt, inget sampel:



Multiplikativ sampelseffekt:



Större modell med fler prediktorer:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &\quad + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 \\ &\quad + \beta_{123} X_1 X_2 X_3 + \varepsilon \end{aligned}$$

med 1.a resp 2.a ordningens samspelseffekter.

Ju högre ordning, desto svårare att tolka!

All samspelseffekter är i praktiken kanske inte nödvändiga (saknar kanske stöd i teori), och saknar stöd i data. Vilka ska man ta med? Testa med partiella F-test!

I Minitab är det enkelt att ta med samspelseffekter, men man måste skapa en kolumn med en ny variabel, tex

$$\text{LET C6} = \text{C1} * \text{C2}$$

och sedan inkludera C6 i regressionskörningen.

Obs! Det är sällan vettigt att ta med en samspelsterm utan att samtidigt inkludera respektive huvudeffekter (se Kleinbaum et al).

Kap 11 forts. Confounding

Confounding, översatt till svenska "förväxling", "bringa i oordning".

Ett starkt regressionssamband mellan två variabler X_1 och Y kan kanske bero på att man inte har *kontrollerat* för en tredje variabel C.

Detta ska synas i att regressionskoefficienten för X_1 kraftigt förändras om vi tar med X_2 jämfört med om vi inte tar med den.

Modell 1, utan att kontrollera för C :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Modell 2, där vi kontrollerar för C :

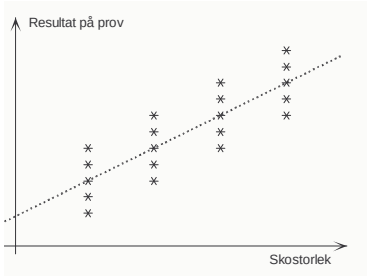
$$Y = \beta_0 + \beta_{1|C} X + \beta_C C + \varepsilon$$

Vi säger att det föreligger "confounding" om det är en *väsentlig skillnad* mellan skattningarna av β_1 och $\beta_{1|C}$, dvs om

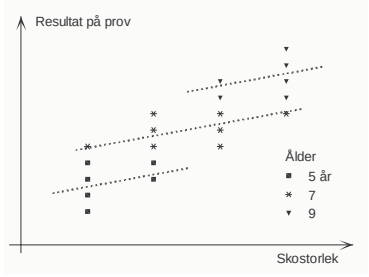
$$\hat{\beta}_1 \neq \hat{\beta}_{1|C}$$

Detta kräver en subjektiv bedömning (se sid. 194-195).

Ex) Sambandet mellan $X =$ Skostorlek och $Y =$ resultat på ett språkprov bland barn, kan kanske se ut som



Men om vi kontrollerar för ålder så har det ursprungliga sambandet (nästan) försvunnit!



Kap 12 Modelldiagnostik, inledning

- Residualanalys: Residualerna skattar ju slumptermerna varpå modellantagandena vilar. Hur kan vi verifiera att antagandena är uppfyllda?
- Outlieranalys: Extrema observationer som antingen uppvisar extremt stora residualer eller har ett starkt inflytande på skattningarna, eller både och.
- Kollinearitet: Hur samvarierar prediktorvariablerna? Är detta ett problem?
- Sakkunskap: Vad är det jag analyserar? Hur har data samlats in? Vad är troliga värden?

Residualer

Definieras

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

Om modellantagandena är uppfyllda bör residualerna uppvisa egenskaper liknande de vi har för slumptermerna, dvs

$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma_\varepsilon^2\right)$$

Vi vet följande:

1. $\bar{e} = \frac{1}{n}\sum e_i = 0$, (garanterat)

2. $S_e^2 = \frac{1}{n-k-1}\sum e_i^2$, (vvr skattning av σ_ε^2 om modellen är rätt)

3. e_i'na är inte oberoende! Varför? Om n är stort är dock inte beroendet starkt och kan bortses ifrån.

Vi definierar följande:

Standardiserade residualer:

$$z_i = \frac{e_i}{S_e}$$

Dessa har en stickprovsvarians $S_z^2 = 1$, vilket underlättar jämförelser mot en standardiserad normal-fördelning.

Leverage-måttet

$$h_i$$

mäter hur viktig observation i är för skattningarna (vi ska återkomma till dessa). Måttet är sådant att $0 \leqslant h_i \leqslant 1$. Beräknas med dator (för svårt för hand).

Studentiserade residualer:

$$r_i = \frac{e_i}{S_e\sqrt{1-h_i}} = \frac{z_i}{\sqrt{1-h_i}}$$

Medelvärdet $\bar{r}$ ligger nära noll och stickprovsvariansen S_r^2 är typiskt något större än 1. Följer approximativt en t-fördelning med $(n-k-1)$ frihetsgrader.

Jackknife residualer:

$$r_{(-i)} = r_i\left(\frac{n-k-1}{n-k-1-r_i^2}\right)^{1/2} = \frac{e_i}{S_{(-i)}\sqrt{1-h_i}}$$

där $S_{(-i)}^2$ är den beräknade residualvariansen (MSE) med den i:te observationen borttagen; baseras mao på $(n-1)$ observationer. Medelvärdet för dessa ligger nära noll och stickprovsvariansen är typiskt något större än 1. Följer approximativt en t-fördelning med $(n-k-2)$ frihetsgrader.

I stora stickprov är standardiserade, studentiserade och jackknife residualer nära nog lika.

Men om det finns observationer som avviker mycket (ger stor residual) eller har ett stort inflytande (ger stort h_i värde) kommer detta att synas tydligare med de två senare.