REGRESSIONSANALYS

Martin Singull

16 april 2018

Innehåll

Kapitel 1

Beroendemått och stokastiska vektorer

Ofta är det intressant att studera flera variabler samtidigt, till exempel kanske villl man undersöka hur en variabel påverkar en annan, eller om det finns något samband mellan variablerna. Det första man vill mäta är då *kovarians* och *korrelation*.

1.1 Beroendemått

Som beroendemått använder man ofta kovarians och korrelation.

Definition 1. *Låt X och Y med väntevärden μ_X respektive μ_Y. Då kallas*

$$\operatorname{cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]$$

för **kovariansen** *mellan X och Y och*

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

för **korrelationen** *mellan X och Y.*

Notera att $\operatorname{cov}(X,Y) = \rho\,\sigma_X\sigma_Y$, där σ_X och σ_Y betecknar standardavvikelserna.

I satsen nedan finns de viktigaste egenskaperna samlade.

Sats 1. *För kovariansen gäller*

(i) $\text{cov}(X, X) = \text{var}(X)$,

(ii) $\text{cov}(X + a, Y + b) = \text{cov}(X, Y)$,

(iii) $\text{cov}(aX, bY) = ab\,\text{cov}(X, Y)$,

(iv) $\text{cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \,\text{cov}(X_i, Y_j)$ *där* $a,\ b,\ a_1, \ldots, a_m,\ b_1, \ldots, b_n$ *är reella konstanter.*

Vidare, för korrelationen gäller att

(v) $|\rho(X, Y)| \leq 1$ *och* $|\rho(X, Y)| = 1$ *om och endast om det finns ett linjärt samband mellan* X *och* Y,

(vi) om X *och* Y *är oberoende så är* $\rho(X, Y) = 0$,

(vii) $\rho(X, Y) = 0$ *medför* **inte** *att* X *och* Y *är oberoende.*

Definition 2. *De s.v.* X *och* Y *kallas okorrelerade om* $\rho(X, Y) = 0$.

Låt $(x_1, y_1), \ldots, (x_n, y_n)$ vara observationer av oberoende och likafördelade stokastiska variabler (X_1, Y_1), $\ldots, (X_n, Y_n)$ med kovarians $\text{cov}(X_j, Y_j) = c$ och korrelation $\rho(X_j, Y_j) = \rho$. Då skattar man kovariansen med

$$\boxed{\hat{c} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}$$

och korrelationen med den empiriska korrelationen

$$\boxed{\hat{\rho} = \frac{\sum_1^n (x_j - \bar{x})(y_j - \bar{y})/(n-1)}{\sqrt{[\sum_1^n (x_j - \bar{x})^2/(n-1)][\sum_1^n (y_j - \bar{y})^2/(n-1)]}}.}$$

Skattningen $\hat{\rho}$ för korrelationen betecknas ofta med r.

Observera att en stark korrelation behöver inte innebära något kausalt samband (orsakssamband) samt att en liten korrelation inte behöver betyda inget samband, see Figur 1.4.

Exempel 1. *Negativ korrelation mellan cigarrettkonsumtion och spädbarnsdödlighet innebär absolut inte att en ökning av cigarrettkonsumtionen ger en minskning av spädbarnsdödligheten.* ■

Exempel 2. *I Figur 1.1 finns samhörande värden på C1: antal lösta radiolicenser/1000 i England, C2: antal personer med mentala defekter per 10 000 invånare. Observationerna är årsvisa värden under en följd av år i radions barndom.* ■

Ofta kan det vara intressant att testa hypotesen

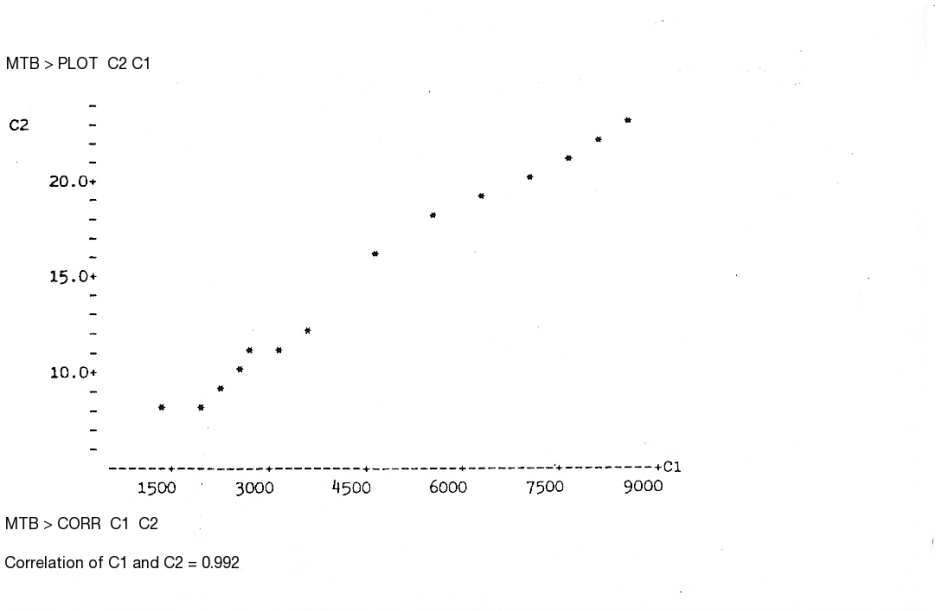
$$H_0 : \rho = 0 \quad \text{mot} \quad H_1 : \rho \neq 0.$$

Detta kan vi göra med teststorheten

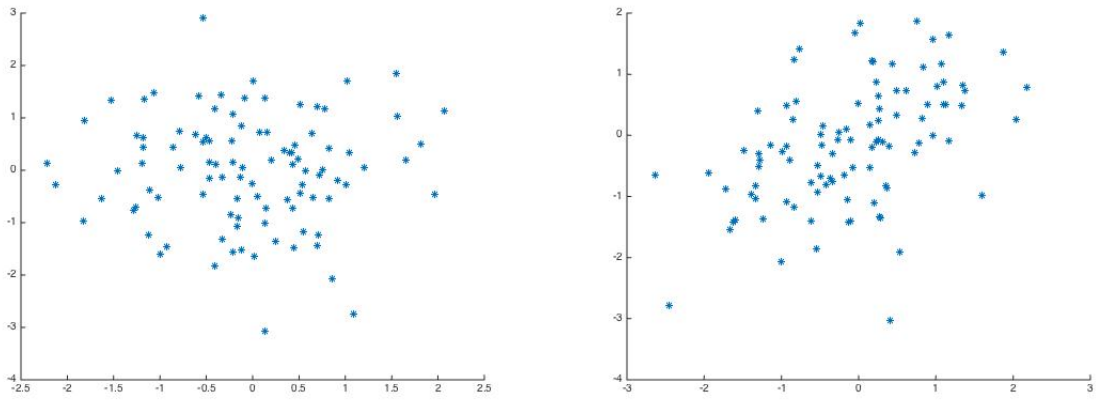
$$u = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}. \tag{1.1}$$

Om H_0 är sann, så gäller vid normalfördelning att den stokastiska variabeln $U \sim t(n-2)$. Man förkastar alltså nollhypotesen på nivå α om

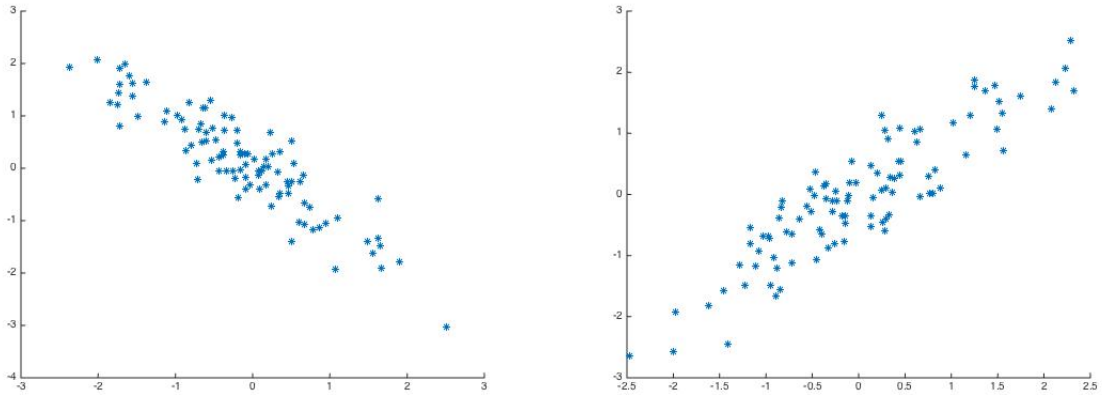
$$|u| > t_{1-\alpha/2}(n-2).$$



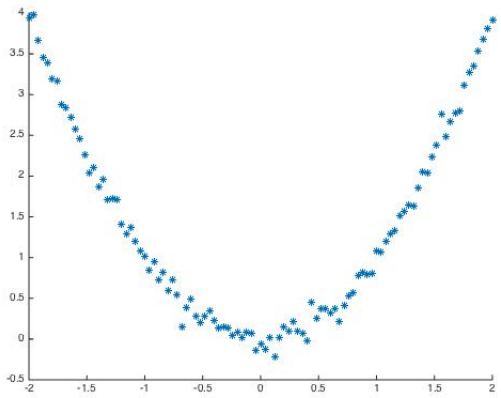
Figur 1.1: C1: antal lösta radiolicenser/1000 i England, C2: antal personer med mentala defekter per 10 000 invånare.



Figur 1.2: $\hat{\rho} = -0.0041$ ($\rho = 0$) till vänster och $\hat{\rho} = 0.4751$ ($\rho = 0.5$) till höger.



Figur 1.3: $\hat{\rho} = -0.9075$ ($\rho = -0.9$) till vänster och $\hat{\rho} = 0.8983$ ($\rho = 0.9$) till höger.



Figur 1.4: $\hat{\rho} = -0.0030$ ($y \sim N(x^2, 0.1)$)

Vi vill nu beräkna variansen för en linjärkombination av stokastiska variabler när kovarianser existerar. Antag att vi har de s.v. $X_1, \ldots, X_n$ med $\text{cov}(X_i, X_j) = \sigma_{ij} \neq 0$ och låt $Y = \sum_{i=1}^{n} a_i X_i$ och vi vill alltså beräkna $\text{var}(Y)$.

Exempel 3. *Vi betraktar först specialfallet när $n = 2$, d.v.s. vi har de s.v. X_1, X_2 med $\text{cov}(X_1, X_2) = \sigma_{12} \neq 0$ och låt $Y = a_1 X_1 + a_2 X_2$. Beräkna $\text{var}(Y)$.*

Lösning: *Låt $\text{E}(X_i) = \mu_i$. Vi har då att*

$$\begin{aligned}
\text{var}(Y) &= \text{E}((a_1 X_1 + a_2 X_2 - (a_1\mu_1 + a_2\mu_2))^2) = \text{E}((a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2))^2) \\
&= \text{E}(a_1^2(X_1 - \mu_1)^2 + a_2^2(X_2 - \mu_2)^2 + 2a_1a_2(X_1 - \mu_1)(X_2 - \mu_2)) \\
&= a_1^2\,\text{E}((X_1 - \mu_1)^2) + a_2^2\,\text{E}((X_2 - \mu_2)^2) + 2a_1a_2\,\text{E}((X_1 - \mu_1)(X_2 - \mu_2)) \\
&= a_1^2\,\text{var}(X_1) + a_2^2\,\text{var}(X_2) + 2a_1a_2\,\text{cov}(X_1, X_2) \\
&= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\sigma_{12}.
\end{aligned}$$

■

Allmän lösning för variansen av $Y = \sum_{i=1}^{n} a_i X_i$ ges av

$$\text{var}(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma_{ij},$$

där $\sigma_{ii} = \sigma_i^2$ och $\sigma_{ij} = \sigma_{ji}$ det vill säga vi har

$$\boxed{\text{var}(Y) = \sum_{i=1}^{n} a_i^2\sigma_i^2 + \sum_{i=1}^{n}\sum_{i<j} 2a_i a_j \sigma_{ij}.}$$

Vi kan beräkna variansen på ett smidigare sätt genom att använda stokastiska vektorer.

1.2 Stokastiska vektorer

En stokastisk vektor definieras som

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} : (n \times 1),$$

där komponenterna X_i är vanliga endimensionella stokastiska variabler.

En stokastisk vektor $\boldsymbol{X}$ har en **väntevärdesvektor**

$$\boldsymbol{\mu} = \text{E}(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} : (n \times 1),$$

där komponenterna $\mu_i = \text{E}(X_i)$ för $i = 1, 2, \ldots, n$. Detta innebär att vi får väntevärdet av en stokastisk matris genom att beräkna väntevärdet av varje element i matrisen.

En stokastisk vektor $\boldsymbol{X}$ har också en **kovariansmatris**

$$\mathbf{C} = \text{cov}(\boldsymbol{X}) = \begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{pmatrix} : (n \times n),$$

där elementen $c_{ij} = \text{cov}(X_i, X_j) = \text{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$ för $i, j = 1, 2, \ldots, n$. En kovariansmatris är alltid *symmetrisk*, $\mathbf{C} = \mathbf{C}'$ (här är $'$ transponat av matrisen), eftersom

$$c_{ij} = \text{cov}(X_i, X_j) = \text{cov}(X_j, X_i) = c_{ji}.$$

Notera att *diagonalelementen* $c_{ii} = \text{var}(X_i)$ för $i = 1, 2, \ldots, n$.

Väntevärdet av en matris är väntevärde för varje element i matrisen, man kan alltså skriva kovariansmatrisen för $\boldsymbol{X}$ som

$$\mathbf{C} = \text{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\right].$$

Exempel 4. *Om X_i och X_j är oberoende för $i \neq j$ med $\text{var}(X_i) = \sigma_i^2$, så är $\text{cov}(X_i, X_j) = 0$ och*

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}.$$

Om dessutom $\text{var}(X_i) = \sigma^2$ för $i = 1, ..., n$, så är

$$\mathbf{C} = \sigma^2 \boldsymbol{I}_n,$$

där

$$\boldsymbol{I}_n = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & & 1 \end{pmatrix} = \text{diag}(1, ..., 1) : n \times n$$

är enhetsmatrisen. ■

Skattning av väntevärdesvektor och kovariansmatris.
Antag att vi har N observationer $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ (vektorer) från någon fördelning med $\text{E}(\boldsymbol{X}_i) = \boldsymbol{\mu}$ och $\text{cov}(\boldsymbol{X}_i) = \mathbf{C}$ för $i = 1, ..., n$. Vi skattar $\boldsymbol{\mu}$ med medelvärdet

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i,$$

och kovariansmatrisen $\mathbf{C}$ med stickprovskovariansmatrisen

$$\mathbf{S} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)' = \frac{1}{N-1} \boldsymbol{X} \left(\boldsymbol{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N'\right) \boldsymbol{X}',$$

där $\boldsymbol{X}$ är observations matrisen $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_N) : n \times N$ och $\mathbf{1}_N$ är en vektor av ettor $\mathbf{1}_N = (1, \ldots, 1)' : (N \times 1)$.

Sats 2. *Låt $\boldsymbol{X} : (n \times 1)$ vara en stokastisk vektor med kovariansmatris $\mathbf{C}_{\boldsymbol{X}}$. Vi definierar en ny stokastisk vektor som*

$$\boldsymbol{Y} = \mathbf{A}\boldsymbol{X} + \boldsymbol{b} : (m \times 1),$$

där $\mathbf{A} : (m \times n)$ är en fix matris och $\boldsymbol{b} : m \times 1$ en fix vektor. Då gäller att $\boldsymbol{Y}$ har väntevärde och kovariansmatris

$$\text{E}(\boldsymbol{Y}) = \mathbf{A}\,\text{E}(\boldsymbol{X}) + \boldsymbol{b},$$
$$\mathbf{C}_{\boldsymbol{Y}} = \mathbf{A}\mathbf{C}_{\boldsymbol{X}}\mathbf{A}'.$$

Bevis. På plats nr. i, i $\boldsymbol{Y}$ har vi $Y_i = \sum_{j=1}^{n} a_{ij}X_j + b_i$ vilket ger

$$\mathrm{E}(Y_i) = \sum_{j=1}^{n} a_{ij}\,\mathrm{E}(X_j) + b_i.$$

Detta innebär att $\mathrm{E}(\boldsymbol{Y}) = \begin{pmatrix} \mathrm{E}(Y_1) \\ \vdots \\ \mathrm{E}(Y_m) \end{pmatrix} = \mathbf{A}\,\mathrm{E}(\boldsymbol{X}) + \boldsymbol{b}$.

Genom att skriva kovariansmatrisen för $\boldsymbol{Y}$ som väntevärdet av en matris, se ovan, får vi

$$\begin{aligned}\mathbf{C}_{\boldsymbol{Y}} &= \mathrm{E}[(\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y}))(\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y}))'] \\ &= \mathrm{E}[(\mathbf{A}\boldsymbol{X} + \boldsymbol{b} - \mathbf{A}\,\mathrm{E}(\boldsymbol{X}) - \boldsymbol{b})(\mathbf{A}\boldsymbol{X} + \boldsymbol{b} - \mathbf{A}\,\mathrm{E}(\boldsymbol{X}) - \boldsymbol{b})'] \\ &= \mathrm{E}[\mathbf{A}(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))'\mathbf{A}'] = \ldots = \mathbf{A}\mathbf{C}_{\boldsymbol{X}}\mathbf{A}'.\end{aligned}$$

Här har vi utnyttjat att $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. □

Specialfall: Variansformel
Låt $X_1, \ldots, X_n$ vara stokastiska variabler. För en linjärkombination av dessa (beroende) stokastiska variabler

$$Y = \sum_{i=1}^{n} a_i X_i = (a_1, \ldots, a_n)\boldsymbol{X} = \mathbf{a}'\boldsymbol{X},$$

där $\boldsymbol{X} = (X_1, \ldots, X_n)'$ och $\boldsymbol{a} = (a_1, \ldots, a_n)'$, gäller att

$$\mathrm{var}(Y) = \sigma_Y^2 = \boldsymbol{a}'\mathbf{C}_{\boldsymbol{X}}\boldsymbol{a} : (1 \times 1).$$

Eftersom $\mathrm{var}(Y) > 0$ får vi också att $\mathbf{C}_{\boldsymbol{X}}$ är positivt definit eller positivt semidefinit.

Exempel 5. *Den stokastiska vektorn* $\begin{pmatrix} X \\ Y \end{pmatrix}$ *har väntevärdesvektor* $\begin{pmatrix} 5 \\ 10 \end{pmatrix}$ *och kovariansmatris* $\begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix}$.

Vi vill förutsäga Y *med hjälp av en prediktor* $\widehat{Y} = a + bX$ *sådan att*

(1) $\mathrm{E}(\widehat{Y}) = \mathrm{E}(Y)$ *och*

(2) $\mathrm{var}(Y - \widehat{Y})$ *är minimal.*

Lösning.

(1) $\mathrm{E}(\widehat{Y}) = \mathrm{E}(a + bX) = a + b\,\mathrm{E}(X) = a + 5b = 10 = \mathrm{E}(Y)$ *ger alltså ekvationen*

$$a + 5b = 10.$$

(2) För variansen gäller att

$$\begin{aligned}\mathrm{var}(Y - \widehat{Y}) &= \mathrm{var}(Y - a - bX) = \mathrm{var}(Y - bX) = \mathrm{var}(\begin{pmatrix} -b & 1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}) \\ &= \begin{pmatrix} -b & 1 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix}\begin{pmatrix} -b \\ 1 \end{pmatrix} = 2b^2 - 6b + 6 = \ldots = 2(b - 1.5)^2 + 1.5,\end{aligned}$$

vilket ger min då $b = 1.5$. *Från ekvationen ovan får vi att* $a = 10 - 5b = 2.5$.
Alltså, välj prediktorn $\widehat{Y} = 2.5 + 1.5X$. ■

Exempel 6. Portföljteori. *Antag att vi har n stycken tillgångar med de stokastiska avkastningarna $\boldsymbol{X} = (X_1, \ldots, X_n)'$, de förväntade avkastningarna $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ och en kovariansmatris $\mathbf{C}$.*

Antag vidare att vi investerar en andelen w_i av vårt totala kapital i tillgång i. Alltså $\sum_i w_i = 1$, där w_i kan vara negativ. Vår portfölj $\boldsymbol{w} = (w_1, \ldots, w_n)'$ har den stokastiska avskastningen

$$R = \boldsymbol{w}'\boldsymbol{X} = \sum_{i=1}^{n} w_i X_i$$

med väntevärde och varians

$$\begin{aligned} \mathrm{E}(R) &= \boldsymbol{w}'\boldsymbol{\mu}, \\ \mathrm{var}(R) &= \boldsymbol{w}'\mathbf{C}\boldsymbol{w}. \end{aligned}$$

Antag nu att vi väljer lika delar av alla tillgångar, dvs. $w_i = \dfrac{1}{n}$ för alla $i = 1, \ldots, n$,

$$\boldsymbol{w} = \frac{1}{n}\mathbf{1}_n = \frac{1}{n}(1, \ldots, 1)' : (n \times 1).$$

1) Om alla avkastningar är oberoende och har variansen $\mathrm{var}(X_i) = \sigma^2$ så får vi kovariansmatrisen $\mathbf{C} = \sigma^2 \boldsymbol{I}_n$. Variansen för portföljen blir nu

$$\mathrm{var}(R) = \boldsymbol{w}'\mathbf{C}\boldsymbol{w} = \frac{1}{n}\mathbf{1}_n' \, \sigma^2 \boldsymbol{I}_n \, \frac{1}{n}\mathbf{1}_n = \frac{\sigma^2}{n^2} \underbrace{\mathbf{1}_n'\mathbf{1}_n}_{=n} = \frac{\sigma^2}{n}.$$

2) Antag att $\mathrm{cov}(X_i, X_j) = 0.3\sigma^2$. Vi har då

$$\mathbf{C} = \sigma^2 \begin{pmatrix} 1 & 0.3 & \cdots & 0.3 \\ 0.3 & 1 & & \vdots \\ \vdots & & \ddots & 0.3 \\ 0.3 & \cdots & 0.3 & 1 \end{pmatrix}$$

och följande varians för portföljen

$$\mathrm{var}(R) = \boldsymbol{w}'\mathbf{C}\boldsymbol{w} = \frac{\sigma^2}{n^2}(1, \ldots, 1) \begin{pmatrix} 1 & 0.3 & \cdots & 0.3 \\ 0.3 & 1 & & \vdots \\ \vdots & & \ddots & 0.3 \\ 0.3 & \cdots & 0.3 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \ldots = \frac{0.7\sigma^2}{n} + 0.3\sigma^2.$$

■

1.3 Flerdimensionell normalfördelning

Först repeterar vi ett par viktiga egenskaper hos stokastiska variabler med endimensionell normalfördelning.

(a) En s.v. $X \sim N(0,1)$, om den har täthetsfunktion $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

(b) Om $Y = \mu + \sigma X$, där $X \sim N(0,1)$, så gäller att $Y \sim N(\mu, \sigma)$.

(c) Om $Y_1, \ldots, Y_n$ är oberoende och normalfördelade, så är även

$$\sum_{i=1}^{n} a_i Y_i + b$$

normalfördelad.

Definitionen av flerdimensionell normalfördelning är en naturlig generalisering av dessa egenskaper.

Definition 3. *En stokastisk vektor* $\boldsymbol{Y} = \begin{pmatrix} Y_1 & \ldots & Y_n \end{pmatrix}' : (n \times 1)$ *har flerdimensionell normalfördelning om*

$$\boldsymbol{Y} = \boldsymbol{\mu} + \mathbf{A}\boldsymbol{X}$$

där $\boldsymbol{\mu} : (n \times 1)$ *är en fix vektor,* $\mathbf{A} : (n \times m)$ *är en fix matris och* $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)'$ *har komponenter* $X_1, \ldots, X_m$*, som är oberoende och* $N(0,1)$*.*

Notera likheten med egenskapen (b) för endimensionell normalfördelning.

Eftersom varje komponent i $\boldsymbol{Y}$ är en konstant plus en linjärkombination av $X_1, \ldots, X_m$, så följer av egenskapen (c) ovan att *varje komponent* Y_i *är normalfördelad.*

Vi ser vidare att $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu} + \mathbf{A0} = \boldsymbol{\mu}$ och att

$$\mathbf{C}_{\boldsymbol{Y}} = \mathbf{A}\mathbf{C}_{\boldsymbol{X}}\mathbf{A}' = \mathbf{A}\boldsymbol{I}_m\mathbf{A}' = \mathbf{A}\mathbf{A}'.$$

Sats 3. *Om* $\boldsymbol{Y}$ *har flerdimensionell normalfördelning med väntevärdesvektor* $\boldsymbol{\mu}$ *och en kovariansmatris* $\mathbf{C}$ *med* $\det \mathbf{C} \neq 0$*, så har* $\boldsymbol{Y}$ *täthetsfunktion*

$$f_{\boldsymbol{Y}}(\boldsymbol{Y}) = f_{\boldsymbol{Y}}(y_1, \ldots, y_n) = \frac{1}{\left(\sqrt{2\pi}\right)^n \sqrt{\det \mathbf{C}}} \, e^{-\frac{1}{2}[(\boldsymbol{Y}-\boldsymbol{\mu})'\mathbf{C}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})]},$$

där $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ *och* $\mathrm{cov}(\boldsymbol{Y}) = \mathbf{C}$*.*

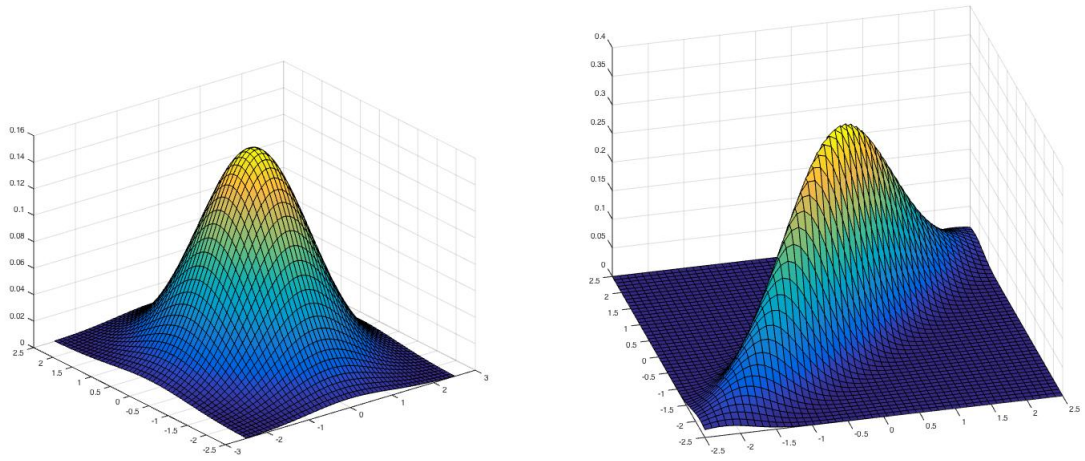
Av den här satsen framgår det att parametrarna för flerdimensionell normalfördelning är väntevärdesvektorn och kovariansmatrisen. Man skriver

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \mathbf{C})$$

eller

$$\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \mathbf{C})$$

om man vill poängtera dimensionen. Om $n = 2$, så säger man att de två s.v. Y_1 och Y_2 har **simultan normalfördelning** om vektorn $\begin{pmatrix} Y_1 & Y_2 \end{pmatrix}'$ är normalfördelad.



Figur 1.5: Täthetsfunktionen för en bivariatnormalfördelning när $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ med $\rho = 0$ till vänster, och $\rho = 0.9$ till höger.

Specialfall: Tvådimensionell normalfördelning
Antag att $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2(\boldsymbol{\mu}, \mathbf{C})$. Då ges täthetsfunktionen av

$$f(y_1, y_2) = \frac{\exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{y_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{y_1 - \mu_1}{\sigma_1} \frac{y_2 - \mu_2}{\sigma_2} + \left(\frac{y_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

där $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ är väntevärdesvektorn, $-1 < \rho < 1$ är korrelationen och kovariansmatrisen ges av

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

med följande villkor $\sigma_1^2 > 0$ och $\sigma_2^2 > 0$. Figur 1.5 visar utseendet för två täthetsfunktioner med olika korrelation.

Sats 4. *Komponenterna i en normalfördelad vektor är oberoende* **om och endast om** *kovariansmatrisen är en diagonalmatris.*

Specialfall: Låt Y_1 och Y_2 vara två simultant normalfördelade s.v. Vi har då

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \quad \text{och} \quad \mathbf{C}_{\boldsymbol{Y}} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix},$$

där $\text{cov}(Y_1, Y_2) = \rho\sigma_1\sigma_2$ och ρ är korrelationen mellan Y_1 och Y_2, σ_1 och σ_2 betecknar standardavvikelserna.

Via satsen ovan har vi att två simultant normalfördelade s.v., Y_1 och Y_2, är oberoende **om och endast om** de är okorrelerade.

Sats 5. *Antag att* $\boldsymbol{d} : (m \times 1)$ *och* $\mathbf{B} : (m \times n)$ *är fixa. Låt*

$$\boldsymbol{W} = \boldsymbol{d} + \mathbf{B}\boldsymbol{Y},$$

där $\boldsymbol{Y} : (n \times 1)$ *har flerdimensionell normalfördelning. Då är även* $\boldsymbol{W} : (m \times 1)$ *normalfördelad.*

Notera att satsen ovan ger att en linjärkombination av beroende normalvariabler, som är komponenter i en normalfördelad vektor, är normalfördelad.

Hur känner man igen normalfördelning?

- Små stickprov: svårt.
- Stora stickprov:
  - endimensionella mätdata - Rita hiostogram och jämför med normalfördelningens täthetsfunktion.
  - tvådimensionella mätdata - Plotta (x_i, y_i). Tendenser till ellipsformat mönster. Histogram kan också göras.

1.3.1 Bevis av en tidigare sats med hjälp av flerdimensionell normalfördelning

Med hjälp av flerdimensionell normalfördelning kan vi nu enkelt bevisa följande sats.

Sats 6. *Om* $X_1, \ldots, X_n$ *är oberoende och* $X_i \sim N(\mu, \sigma)$*, så gäller att*

(a) $\sum_1^n \left(\dfrac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$

(b) $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum_1^n \left(X_i - \bar{X}\right)^2}{\sigma^2} \sim \chi^2(n-1)$

(c) $\bar{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$

(d) $\bar{X}$ *och* S^2 *är oberoende stokastiska variabler.*

Bevis. (c) har vi redan visat på Fö1 och (a) ser man att det är en summa av kvadrater på oberoende $N(0,1)$-variabler.

(b) $\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 = \sum_{i=1}^n \left(Z_i - \bar{Z}\right)^2 = \boldsymbol{Z}' \underbrace{\left(\boldsymbol{I}_n - \dfrac{1}{n}\mathbf{1}_n\mathbf{1}_n' \right)}_{=\mathbf{C}} \boldsymbol{Z},$

med $\boldsymbol{Z} = (Z_1, \ldots, Z_n)' \sim N_n\left(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n\right)$, där $Z_i = X_i - \mu$.

Matrisen $\mathbf{C} : n \times n$ är symmetrisk och idempotent (projektions matris), d.v.s. $\mathbf{C}^2 = \mathbf{C}$ med rang $n-1$ och har således $n-1$ egenvärden som är 1 och ett egenvärde som är 0. Vi har nu spektraluppdelningen av $\mathbf{C}$ som

$$\mathbf{C} = \mathbf{Q}\mathbf{D}\mathbf{Q}',$$

där $\mathbf{D}$ är en diagonalmatris med egenvärdena på diagonalen, d.v.s., $\mathbf{D} = diag(1, \ldots, 1, 0)$ och $\mathbf{Q}$ är en

ortonormerad matris, d.v.s. $\mathbf{Q}'\mathbf{Q} = \boldsymbol{I}$. Vi har nu

$$\boldsymbol{Z}'\mathbf{C}\boldsymbol{Z} = (\boldsymbol{Z}'\mathbf{Q})\mathbf{D}(\mathbf{Q}'\boldsymbol{Z}) = \boldsymbol{Y}' \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & 1 & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} \boldsymbol{Y},$$

där $\boldsymbol{Y} = \mathbf{Q}'\boldsymbol{Z} \sim N_n\big(\mathbf{0}, \underbrace{\sigma^2\mathbf{Q}'\mathbf{Q}}_{=\sigma^2\boldsymbol{I}}\big)$. Nu gäller att

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_1^n \left(X_i - \bar{X}\right)^2}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{Z}'\mathbf{C}\boldsymbol{Z} = \frac{1}{\sigma^2}\boldsymbol{Y}' \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & 1 & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} \boldsymbol{Y} = \sum_{i=1}^{n-1} \left(\frac{Y_i}{\sigma}\right)^2 \sim \chi^2(n-1),$$

eftersom $\dfrac{Y_i}{\sigma} \sim N(0,1)$ $\big($då $Y_i \sim N(0,\sigma)\big)$.

(d) Vi har att $\bar{X} = \frac{1}{n}\mathbf{1}_n'\boldsymbol{X}$. Man kan visa att $\bar{X}$ och S^2 är oberoende eftersom

$$\left(\frac{1}{n}\mathbf{1}'\right)\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right) = \mathbf{0}'.$$

□

Kapitel 2

Regressionsanalys

Regressionsanalys används för att studera sambandet mellan en boroende kvantitativ variabel y och en eller flera variabler $x_1, x_2, \ldots, x_k$. Variablerna $x_1, x_2, \ldots, x_k$ brukar något felaktigt kallas "oberoende" variabler. Det betyder inte att de måste vara oberoende av varandra bara att är y som beror på x-variablerna och inte tvärtom. Syftet med regressionsanalysen är att finna modeller som är enkla men ändå passar bra med sina observationer. Modellerna klan vara linjära eller icke-linjära. Nedan börjar vi först med enkel linjär regression innan vi beskriver det mer allmänt med multipel linjär regression.

2.1 Enkel linjär regression

Vi börjar med ett inledande exempel.

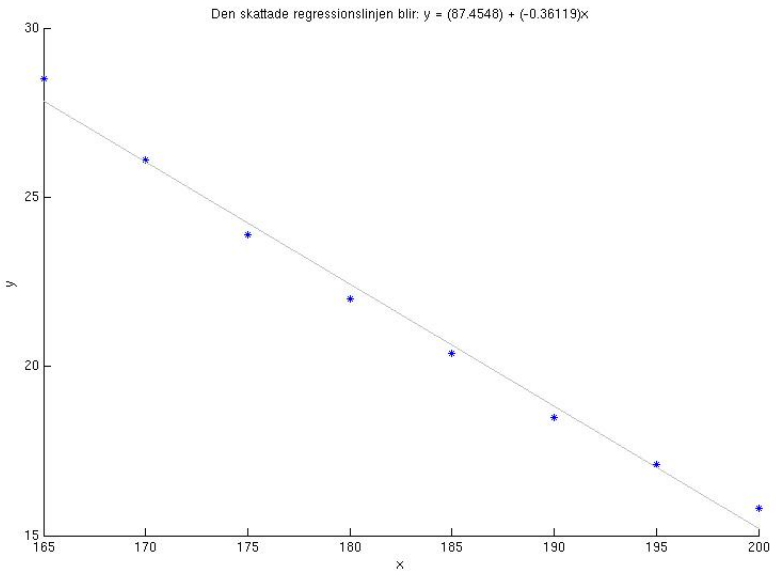
Exempel 7. *Viskositeten hos motorolja avtar med temperaturen. Samhörande värden på åtta viskositeter* $((lb)(sec)/(in.)^2)$ *och temperaturer* $(^\circ F)$ *har mätts upp*

Temp. (x_j)*:*	*165*	*170*	*175*	*180*	*185*	*190*	*195*	*200*
Visk. (y_j)*:*	*28.5*	*26.1*	*23.9*	*22*	*20.4*	*18.5*	*17.1*	*15.8*

Vi kan nu plotta y-värdena mot x-värdena.

```
%MATLAB
y = [28.5 26.1 23.9 22 20.4 18.5 17.1 15.8]';
x = [165 170 175 180 185 190 195 200]';

scatter(x,y,'b*')
xlabel('x'), ylabel('y')
hold on
lsline
```

Vidare så har vi den empiriska korrelationen



Figur 2.1:

```
>> r = corr(x,y)

r =

   -0.9955
```

Det vill säga $\hat{\rho} = -0.9955$ *vilken till beloppet är väldigt stor. men, för sakens skull så testar vi hypotesen* $H_0 : \rho = 0$ *mot* $H_1 : \rho \neq 0$ *på nivån 5%, med testorheten som ges i* (1.1)*. Vi har alltså*

$$u = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} = -25.73$$

och eftersom $|u| > t_{0.975}(6) = 2.45$ *så kan vi förkasta* H_0*, det vill säga korrelationen är med stor sannolikhet skild från noll.*

Från Figur 2.1 och korrelationen ovan så verkar det rimligt med ett approximativt linjärt samband. Men frågan kvarstår om hur vi hittar den räta linje som passar "bäst" till de observerade värdena. ■

I enkel linjär regression så har vi värdepar (x_j, y_j), där y_j är observation av den stokastiska variabeln

$$Y_j = \mu_j + \varepsilon_j = \beta_0 + \beta_1 x_j + \varepsilon_j,$$

för $j = 1, \ldots, n$, där

$$\mu_j = \beta_0 + \beta_1 x_j$$

och $x_1, \ldots, x_n$ är fixa tal medan $\varepsilon_1, \ldots, \varepsilon_n$ är oberoende stokastiska variabler med $E(\varepsilon_j) = 0$ och $Var(\varepsilon_j) = \sigma^2$. Vi ska nu gå över till det allmänna fallet - *multipel linjär regression.*

2.2 Multipel linjär regression

I det allmänna fallet vill man förklara variationerna hos en responsvariabel y med hjälp av variationerna hos förklaringsvariabler $x_1, \ldots, x_k$. Man formulerar ett "teoretiskt" linjärt samband

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

För att skatta β-koefficienterna i det linjära sambandet behöver vi observerade värden

$$\begin{gathered}(x_{11}, x_{12}, \ldots, x_{1k}, y_1) \\ \vdots \\ (x_{n1}, x_{n2}, \ldots, x_{nk}, y_n)\end{gathered}$$

I allmänhet upptäcker man sedan att sambandet mellan de observerade y- och x-värdena inte är exakt linjärt och behöver då också analysera felen som uppstår. Även i det allmänna fallet gör man en modell som innebär att avvikelserna från det linjära sambandet betraktas som slumpvariabler.

Modell: Varje y_j är observation av

$$Y_j = \mu_j + \varepsilon_j = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_k x_{jk} + \varepsilon_j, \tag{2.1}$$

för $j = 1, \ldots, n$, där

$$\mu_j = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_k x_{jk} \tag{2.2}$$

och $x_{j1}, \ldots, x_{jk}$ är fixa kända tal, $\beta_0, \ldots, \beta_k$ är okända parametrar och där

$$\varepsilon_1, \ldots, \varepsilon_n \text{ är oberoende och } N(0, \sigma).$$

Modellen ger att de s.v. $Y_1, \ldots, Y_n$ är oberoende stokastiska variabler sådana att

$$Y_j \sim N(\mu_j, \sigma),$$

där μ_j ges av (2.2). Vi skriver modellen med matriser i stället

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} \\ \beta_0 + \beta_1 x_{21} + \ldots + \beta_k x_{2k} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Alltså, vi har

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{=\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{=\boldsymbol{\varepsilon}}$$

eller kortare

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Anm. Om man gör regressionsanalys **utan konstantterm** β_0 så blir formlerna likadana men designmatrisen $\boldsymbol{X}$ blir annorlunda.

Vidare gäller att $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ och $\mathbf{C}_{\boldsymbol{Y}} = \sigma^2 \boldsymbol{I}_n$ eftersom ε_i är oberoende $N(0, \sigma)$, $i = 1, \ldots, n$, det vill säga $\mathbf{C}_{\boldsymbol{\varepsilon}} = \sigma^2 \boldsymbol{I}_n$. Med hjälp av flerdimensionell normalfördelning kan vi skriva den linjära modellen som

$$\boxed{\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n).} \tag{2.3}$$

Minsta-kvadrat-skattningen (MK-skattningen) av parametrarna $\boldsymbol{\beta}$ ges av att minimera funktion

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{j=1}^{n} (y_j - (\beta_0 + \beta_1 x_{j1} + \ldots + \beta_k x_{jk}))^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned}$$

Sats 7. *Under förutsättningarna i modellen* (2.3) *och om* $\det\left(\boldsymbol{X}'\boldsymbol{X}\right) \neq 0$, *så gäller att MK-skattningen är*

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Låt de skattade y-värdena ("fitted values") vara $\hat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}$, *där* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. *Residualerna ges av*

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y},$$

vilka uppfyller $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$ *and* $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$.

Bevis. Matrisen $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ uppfyller

1. $(\boldsymbol{I} - \boldsymbol{H})' = \boldsymbol{I} - \boldsymbol{H}$ (symmetrisk)
2. $(\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$ (idempotent)
3. $\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$.

Vidare gäller nu att $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{0}$ och $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = 0$. Vi ska nu minimera funktion $Q(\boldsymbol{\beta})$. Funktionen kan skrivas som

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \end{aligned}$$

eftersom $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{X} = \hat{\boldsymbol{\varepsilon}}'\boldsymbol{X} = \boldsymbol{0}'$. Första termen i $Q(\boldsymbol{\beta})$ beror inte på $\boldsymbol{\beta}$ och för den andra termen gäller att $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \geq 0$ med likhet för $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. Därmed har vi visat att $Q(\boldsymbol{\beta}) \geq Q(\widehat{\boldsymbol{\beta}})$ och eftersom $\det\left(\boldsymbol{X}'\boldsymbol{X}\right) \neq 0$ har vi likhet om och endast om $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. □

Sats 8. *Under förutsättningarna i modellen ovan så gäller att den stokastiska vektorn*

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right).$$

Bevis. Eftersom $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ så ser vi att $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ är en linjärkombination av normalfördelning och således är även $\widehat{\boldsymbol{\beta}}$ normalfördelad. Vidare gäller att

$$\mathrm{E}(\boldsymbol{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{E}(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

och

$$\begin{aligned} \mathbf{C}_{\widehat{\boldsymbol{\beta}}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{C}_{\mathbf{Y}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned}$$

Alltså, vi har att $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$. □

2.3 Variansanalys

Variansanalys handlar om att studera kvadratsumman som beskriver den totala variationen i y-värdena och dela upp den nya kvadratsummor som mäter olika saker.

Låt

$$\hat{\mu}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_k x_{jk},$$

det vill säga skattningen av $\mathrm{E}(Y_j)$ enligt regressionsmodellen. Man kan visa att

$$\underbrace{\sum_{j=1}^{n}(y_j - \bar{y})^2}_{=SS_{TOT}} = \underbrace{\sum_{j=1}^{n}(\hat{\mu}_j - \bar{y})^2}_{=SS_R} + \underbrace{\sum_{j=1}^{n}(y_j - \hat{\mu}_j)^2}_{=SS_E},$$

det vill säga

$$\boxed{SS_{TOT} = SS_R + SS_E,}$$

där

$$SS_{TOT} = \sum_{j=1}^{n}(y_j - \bar{y})^2,$$
$$SS_R = \sum_{j=1}^{n}(\hat{\mu}_j - \bar{y})^2,$$
$$SS_E = \sum_{j=1}^{n}(y_j - \hat{\mu}_j)^2.$$

Här beskriver

- SS_{TOT} (även kallad Q_{TOT}) - totala variationen hos y-värdena.
- SS_R (även kallad Q_{REGR}) - variation hos y-värdena som förklaras av $x_1, \ldots, x_k$. Man kan visa att SS_R är en positivt definit kvadratisk form i $\hat{\beta}_1, \ldots, \hat{\beta}_k$.
- SS_E (även kallad Q_{RES}) - variation hos y-värdena som vi inte lyckats förklara via regressionsmodellen.

Exempel 8. *För att beskriva de tre olika kvadratsummorna kan följande modeller och plottar vara till hjälp. Betrakta först modellen $Y_j = \mu + \varepsilon_j$ där vi skattar $\mu = \mathrm{E}(Y_j)$ med $\bar{y}$. Den totala variationen hos datan ges av*

$$SS_{TOT} = \sum_{j=1}^{n}(y_j - \bar{y})^2$$

och kan beskrivas med Figur 8.

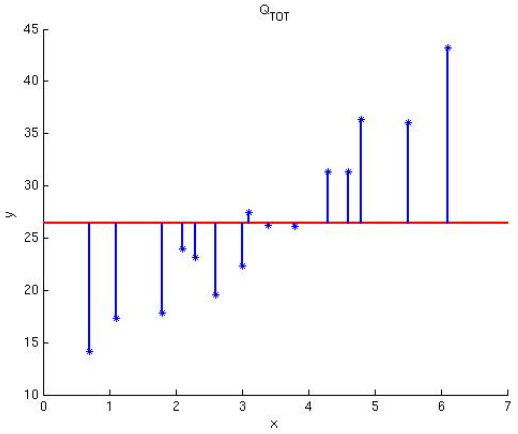
Modellen som tar hänsyn till den linjära regressionen ges av $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ där vi skattar β_0 och β_1 som i Sats 7. Variationen som hänger samman med regressionsmodellen ges av

$$SS_R = \sum_{j=1}^{n}\left(\hat{\mu}_j - \bar{y}\right)^2 = \sum_{j=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_j - \bar{y}\right)^2$$

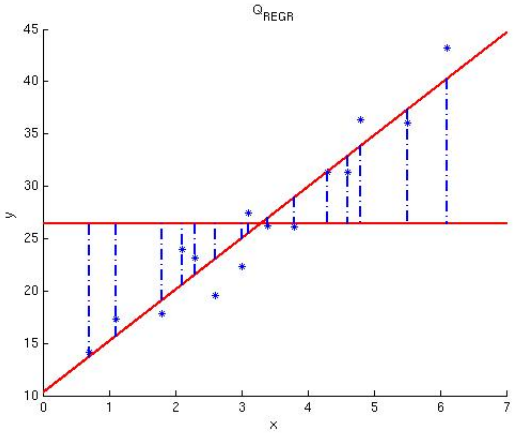
och kan beskrivas med Figur 2.3.

Sist har vi variationen som inte den linjära modellen kan förklara. Den variationen ges av

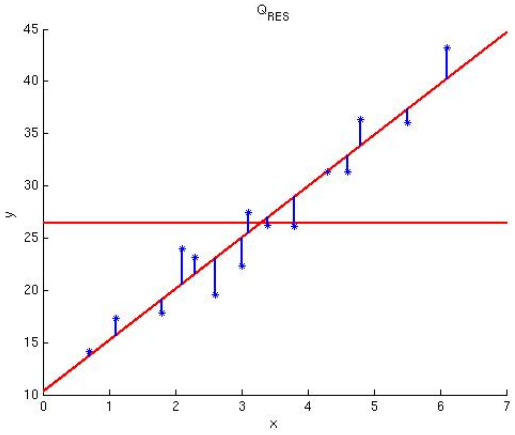
$$SS_E = \sum_{j=1}^{n}\left(y_j - \hat{\mu}_j\right)^2 = \sum_{j=1}^{n}\left(y_j - (\hat{\beta}_0 + \hat{\beta}_1 x_j)\right)^2$$



Figur 2.2: SS_{TOT}



Figur 2.3: SS_R



Figur 2.4: SS_E

och kan beskrivas med Figur 2.4.

■

Förklaringsgraden: $R^2 = \dfrac{SS_R}{SS_{TOT}}$

Notera att $R^2 = \dfrac{SS_{TOT} - SS_E}{SS_{TOT}} = 1 - \dfrac{SS_E}{SS_{TOT}}$.

Ibland studerar man i stället $R^2_{ADJ} = 1 - \dfrac{SS_E/(n-k-1)}{SS_{TOT}/(n-1)}$ som "straffar" om k är stort i förhållande till n.

Exempel 9. *I Exempel 7 har vi samhörande värden på viskositet* $((lb)(sec)/(in.)^2)$ *och temperatur* $(^\circ F)$

Temp. (x_j)*:*	*165*	*170*	*175*	*180*	*185*	*190*	*195*	*200*
Visk. (y_j)*:*	*28.5*	*26.1*	*23.9*	*22*	*20.4*	*18.5*	*17.1*	*15.8*

Detta ger modellen

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{15} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 165 \\ 1 & 170 \\ \vdots & \vdots \\ 1 & 200 \end{pmatrix}}_{=\boldsymbol{X}} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{15} \end{pmatrix}$$

och en observerad vektor $\mathbf{y} = \begin{pmatrix} 28.5 & 26.1 & \dots & 15.8 \end{pmatrix}'$. *Vidare har vi*

$$\boldsymbol{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 165 & 170 & \cdots & 200 \end{pmatrix} \begin{pmatrix} 28.5 \\ \vdots \\ 15.8 \end{pmatrix} = \begin{pmatrix} 172.3 \\ 31065.5 \end{pmatrix},$$

och

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 165 & 170 & \cdots & 200 \end{pmatrix} \begin{pmatrix} 1 & 165 \\ 1 & 170 \\ \vdots & \vdots \\ 1 & 200 \end{pmatrix} = \begin{pmatrix} 8 & 1460 \\ 1460 & 267500 \end{pmatrix},$$

samt

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{8400} \begin{pmatrix} 267500 & -1460 \\ -1460 & 8 \end{pmatrix} = \begin{pmatrix} 31.84523810 & -0.17380952 \\ -0.17380952 & 0.00095238 \end{pmatrix}$$

och skattningarna

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} 87.4549 \\ -0.3612 \end{pmatrix}.$$

Det vill säga, vi har $\hat{\beta}_0 = 87.4549$ *och* $\hat{\beta}_1 = -0.3612$ *samt*

$$SS_E = \sum_{j=1}^{8} (y_j - (87.4549 - 0.3612x_j))^2 = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 1.2373.$$

Från satsen ovan har vi att de s.v.

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma\sqrt{31.84523810}\right) \quad \text{och} \quad \hat{\beta}_1 \sim N\left(\beta_1, \sigma\sqrt{0.00095238}\right).$$

Vi kan också skatta regressionsparametrarna med hjälp av MATLAB och kommandot `regstats`*. Använd hjälpfunktionen för att se hur kommandot fungerar i detalj.*

```
regr = regstats(y,x,'linear','all');
```

Vi får då följande resultat.

```
>> regr.tstat.beta

ans =

   87.4548
   -0.3612

>> regr.fstat.sse

ans =

    1.2373
```

■

Sats 9 (Huvudsats)**.** *Under förutsättningarna i modellen ovan gäller gäller att*

a) den s.v. $\dfrac{SS_E}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{j=1}^{n}(Y_j - \hat{\mu}_j)^2 \sim \chi^2(n-k-1)$.

b) om $\beta_1 = \ldots = \beta_k = 0$ *så är den s.v.*
$\dfrac{SS_R}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n}(\hat{\mu}_j - \bar{Y})^2 \sim \chi^2(k)$.

c) den s.v. SS_E *är oberoende av den s.v.* SS_R *och av den stokastiska vektorn* $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

Konsekvenser av satsen:

a) ger $\boxed{\sigma^2\text{-skattningen } s^2 = \dfrac{SS_E}{n-k-1},}$

som är väntevärdesriktig eftersom

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{\sigma^2}{n-k-1}\frac{SS_E}{\sigma^2}\right) = \frac{\sigma^2}{n-k-1}\,\mathrm{E}\left(\frac{SS_E}{\sigma^2}\right) = \frac{\sigma^2}{n-k-1}(n-k-1) = \sigma^2.$$

c) Används vid konstruktion av konfidensintervall.

b) Används då man prövar

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{(alla förklaringsvariablerna är meningslösa)}$$

mot

$$H_1 : \text{minst ett } \beta_i \neq 0 \quad \text{(minst en förklaringsvariabel gör nytta).}$$

$\boxed{\text{Teststorhet: } v = \dfrac{SS_R/k}{SS_E/(n-k-1)}.}$

Den s.v. $V \sim F(k, n-k-1)$ om H_0 är sann och avvikelser från H_0 ger stora värden på SS_{REGR}. Alltså förkastas H_0 om v är **stor** det vill säga om $v > c = F_{1-\alpha}(k, n-k-1)$.

Exempel 10. *Fortsättning på Exempel 9. Vi kan nu skatta variansen* σ^2 *med*

$$s^2 = \frac{SS_E}{n-k-1} = \frac{1.2373}{6} = 0.2062$$

med 6 frihetsgrader.

Vi vill nu testa hypotesen

$$H_0 : \beta_1 = 0 \quad mot \quad H_1 : \beta_1 \neq 0$$

och kan göra det med hjälpvariabeln

$$v = \frac{SS_R}{SS_E/(n-2)} = 6.64,$$

där $SS_R = SS_{TOT} - SS_E = 136.9815$ *och* $SS_{TOT} = \sum_{j=1}^{8}(y_j - \bar{y})^2 = 138.2188$ *samt* $\bar{y} = 21.5375$. *Förkasta* H_0 *om* $v > c = F_{0.95}(1,6) = 5.99$. *Alltså, förkasta* H_0. *Förklaringsvariabeln* x *gör nytta med stor sannolikhet.*

Med MATLAB har vi

```
>> regr.fstat

ans =

     sse: 1.237261904761912
     dfe: 6
     dfr: 1
     ssr: 1.369814880952378e+02
       f: 6.642804772442936e+02
    pval: 2.249040699297267e-07
```

Alltså, vi ser också att P-värdet för testet ovan är mindre än 5%.

Förklaringsgraden är nu lätt att räkna ut som

$$R^2 = \frac{SS_R}{SS_{TOT}} = \frac{136.9815}{138.2188} = 0.9910,$$

vilket tyder på väldigt bra anpassning. Om vi gör det i MATLAB så får vi

```
>> regr.rsquare

ans =

    0.9910
```

■

För modellen

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \tag{2.4}$$

gäller sammanfattningsvis att

(i) att den stokastiska vektorn $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$,

(ii) att de stokastiska variablerna $\widehat{\boldsymbol{\beta}}$ och SS_E är oberoende,

(iii) att $\dfrac{SS_E}{\sigma^2} = \dfrac{(n-k-1)S^2}{\sigma^2} \sim \chi^2(n-k-1)$.

Vi inför nu beteckningarna

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} h_{00} & h_{01} & \dots & h_{0k} \\ h_{10} & h_{11} & \dots & h_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ h_{k0} & h_{k1} & \dots & h_{kk} \end{pmatrix}$$

(i) ger att den s.v. $\hat{\beta}_i \sim N(\beta_i, \sigma\sqrt{h_{ii}})$. Alltså

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{h_{ii}}} \sim N(0,1).$$

Genom att utnyttja (ii) och (iii) samt Gossets sats får vi

$$\frac{(\hat{\beta}_i - \beta_i)/(\sigma\sqrt{h_{ii}})}{\sqrt{\left[(n-k-1)\frac{S^2}{\sigma^2}\right]/(n-k-1)}} \sim t(n-k-1)$$

d.v.s.

$$\frac{\hat{\beta}_i - \beta_i}{S\sqrt{h_{ii}}} \sim t(n-k-1)$$

och detta är vår hjälpvariabel för konstruktion av I_{β_i} som ges av

$$\boxed{I_{\beta_i} = \left(\hat{\beta}_i \mp t_{1-\alpha/2}(n-k-1)\; s\; \sqrt{h_{ii}}\right)}$$

med signifikansnivån α. Om vi direkt vill testa hypotesen

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

så använder vi teststorheten

$$\boxed{t = \frac{\hat{\beta}_i}{S\sqrt{h_{ii}}}.}$$

Om H_0 är sann, så gäller att den stokastiska variabeln $T \sim t(n-k-1)$. Man förkastar alltså nollhypotesen på nivå α om

$$|t| > t_{1-\alpha/2}(n-k-1).$$

Exempel 11. *Fortsättning på Exempel 10. Vi vill nu testa hypotesen*

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

på två nya sätt.

1. *Bilda konfidensintervall för β_1. Hjälpvariabeln ges nu av*

$$\frac{\hat{\beta}_1 - \beta_1}{S\sqrt{h_{11}}} \sim t(6)$$

vilken ger intervallet

$$I_{\beta_1} = \left(\hat{\beta}_1 \mp ts\sqrt{h_{11}}\right),$$

där $\hat{\beta}_1 = -0.3612$, $t = t_{0.975}(6) = 2.45$, $s = \sqrt{0.2062}$ och $h_{11} = 0.00095238$. Vi får alltså intervallet

$$I_{\beta_1} = (-0.40\ ,\ -0.33)\,.$$

Då noll inte finns i intervallet så kan vi förkasta H_0. Förklaringsvariabeln x gör nytta med stor sannolikhet.

Vi kan bilda designmatrisen $\boldsymbol{X}$ *i MATLAB och sedan använda en färdig rutin för att beräkna konfidensintervallet.*

```
n = length(x);
X = [ones(n,1) x];
```

Nu när vi har bildat designmatrisen $\boldsymbol{X}$ *kan vi också använda kommandot* `regress`*.*

```
>> [bhat,bint] = regress(y,X)

bhat =

   87.4548
   -0.3612


bint =

   81.1844   93.7252
   -0.3955   -0.3269
```

Där vi ser att intervallet för β_1 *ges av* $I_{\beta_1} = (-0.40\ ,\ -0.33)$ *.*

2. *Vi kan också testa hypotesen* H_0 *med teststorheten*

$$t = \frac{\hat{\beta}_1}{s\sqrt{h_{11}}} = -25.77,$$

där $\hat{\beta}_1 = -0.3612$*,* $s = \sqrt{0.2062}$ *och* $h_{11} = 0.00095238$*. Om* H_0 *är sann är den stokastiska variabeln* $T \sim t(6)$ *och vi ska förkasta* H_0 *om* $|t| > c = t_{0.975}(6) = 2.45$*. Alltså, vi kan förkasta* H_0*. Förklaringsvariabeln* x *gör nytta med stor sannolikhet.*

Återigen med MATLAB har vi

```
>> regr.tstat

ans =

    beta: [2x1 double]
      se: [2x1 double]
       t: [2x1 double]
    pval: [2x1 double]
     dfe: 6
```

där

```
>> regr.tstat.se

ans =

    2.5626
    0.0140
```

```
>> regr.tstat.t

ans =

   34.1276
  -25.7736

>> regr.tstat.pval

ans =

   1.0e-06 *

    0.0422
    0.2249
```

Medelfelet $d(\hat{\beta}_1) = s\sqrt{h_{11}} = \sqrt{0.2062 \cdot 0.00095238} = 0.0140$ *finns i* `se` *i MATLAB-utskriften.* ■

2.4 Konfidens- och prediktionsintervall

Antag att vi har en observation $\boldsymbol{y}$ som är en observation från $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, där $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$, $\boldsymbol{X} : n \times (k+1)$ är en känd designmatris och $\boldsymbol{\beta} : (k+1) \times 1$ är okända regressionsparametrar. Vi är intresserade av en ny ännu inte observerad stokastisk variabel

$$Y_0 = \underbrace{\beta_0 + u_1\beta_1 + \ldots + u_k\beta_k}_{=\mu_0} + \varepsilon_0 = \mu_0 + \varepsilon_0,$$

där $u_1, \ldots, u_k$ är de aktuella värdena på förklaringsvariablerna dvs. kända tal, där ε_0 är oberoende av $\varepsilon_1, \ldots, \varepsilon_n$ och

$$\varepsilon_0 \sim N(0, \sigma).$$

Vi inför vektorn

$$\boldsymbol{u}' = (1 \; u_1 \; \ldots \; u_k)$$

och kan då skriva

$$Y_0 = \boldsymbol{u}'\boldsymbol{\beta} + \varepsilon_0 = \mu_0 + \varepsilon_0,$$

där $\mu_0 = \boldsymbol{u}'\boldsymbol{\beta}$. Vi skattar μ_0 med $\hat{\mu}_0 = \boldsymbol{u}'\widehat{\boldsymbol{\beta}}$. Väntevärdet av $\hat{\mu}_0$ är

$$\mathrm{E}(\hat{\mu}_0) = \mathrm{E}\left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{u}'\,\mathrm{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{u}'\boldsymbol{\beta} = \mu_0$$

och variansen ges av

$$\mathrm{var}(\hat{\mu}_0) = \mathrm{var}\left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{u}'\boldsymbol{C}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{u} = \sigma^2\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}.$$

Vidare, eftersom den s.v. $\hat{\mu}_0$ är en linjär transformation av en normalfördelad vektor så blir $\hat{\mu}_0$ normalfördelad, det vill säga

$$\hat{\mu}_0 = \boldsymbol{u}'\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{u}'\boldsymbol{\beta}, \sigma\sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}\right).$$

Genom att standardisera variabeln ovan och skatta σ^2 med $s^2 = \dfrac{SS_E}{n-k-1}$ får vi hjälpvariabeln

$$\frac{\hat{\mu}_0 - \mu_0}{S\sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}} = \frac{\boldsymbol{u}'\widehat{\boldsymbol{\beta}} - \boldsymbol{u}'\boldsymbol{\beta}}{S\sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}} \sim t(n-k-1)$$

som vi använder för konstruktion av konfidensintervall för $E(Y_0) = \mu_0 = \boldsymbol{u}'\boldsymbol{\beta}$. Observera att frihetsgraden i t-fördelningen kommer från σ^2-skattningens χ^2-variabel. Konfidensintervallet ges nu av

$$I_{\mu_0} = \left(\hat{\mu}_0 \mp t_{1-\alpha/2}(n-k-1)\ s\ \sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}} \right)$$

eller

$$\boxed{I_{\boldsymbol{u}'\boldsymbol{\beta}} = \left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}} \mp t_{1-\alpha/2}(n-k-1)\sqrt{\boldsymbol{u}'\widehat{\boldsymbol{C}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{u}} \right)},$$

där $\widehat{\boldsymbol{C}}_{\widehat{\boldsymbol{\beta}}} = s^2\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}$. Notera att väntevärdet μ_0 är ett "teoretiskt" värde som beskriver vad man får i genomsnitt om man gör många mätningar med samma värde på förklaringsvariablerna.

Ofta är man i stället intresserad av vad som kan hända i en **enskild** mätning det vill säga vilka värden som är tänkbara för den stokastiska variabeln Y_0. Då konstruerar man ett så kallat **prediktionsintervall**. Knepet är att utnyttja den stokastiska variabeln

$$Y_0 - \hat{\mu}_0 = Y_0 - \boldsymbol{u}'\widehat{\boldsymbol{\beta}} \sim N\left(0, \sigma\sqrt{1 + \boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}\right) \tag{2.5}$$

eftersom de stokastiska variablerna $Y_0 \sim N(\mu_0, \sigma)$ och

$$\hat{\mu}_0 = \boldsymbol{u}'\widehat{\boldsymbol{\beta}} \sim N\left(\mu_0, \sigma\sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}\right)$$

är oberoende och

$$\mathrm{var}(Y_0 - \hat{\mu}_0) = \mathrm{var}(Y_0) + \mathrm{var}(\hat{\mu}_0).$$

Genom att standardisera (2.5) och skattta σ^2 med s^2 får vi hjälpvariabeln för konstruktion av prediktionsintervall för Y_0 som

$$\frac{Y_0 - \boldsymbol{u}'\widehat{\boldsymbol{\beta}}}{S\sqrt{1 + \boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}} \sim t(n-k-1).$$

Prediktionsintervallet ges nu av

$$\boxed{I_{Y_0} = \left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}} \mp t_{1-\alpha/2}(n-k-1)\sqrt{s^2 + \boldsymbol{u}'\widehat{\boldsymbol{C}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{u}} \right).}$$

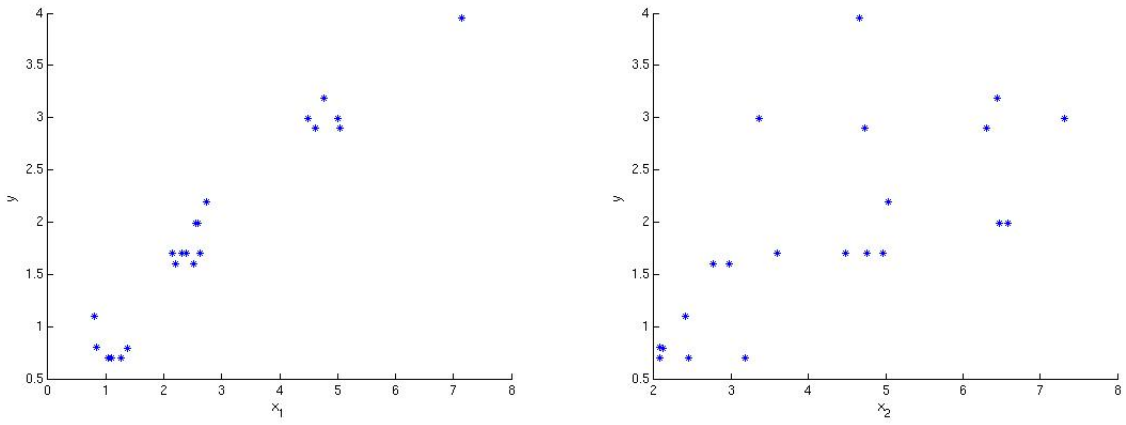
Exempel 12. *Vid ett tillfälle skulle en ny hamburgerrestaurangkedja etablera sig i USA och man var intresserad av att tillämpa samma prispolitik som de gamla hamburgerkedjorna. Man samlade därför in uppgifter om*

$$\begin{aligned} y &= \textit{priset på en hamburgare (enhet: dollar)} \\ x_1 &= \textit{vikten av det tillagade köttet (enhet: oz.)} \\ x_2 &= \textit{vikten av brödet och andra tillbehör (enhet: oz.)} \end{aligned}$$

Resultat:

```
y = [0.79 0.8 1.6 0.7 0.7 0.7 1.1 1.7 1.7 1.99 3.19 2.99 2.19 1.7 1.99 ...
       2.99 1.6 1.7 2.9 2.9 3.95]';
x1 = [1.38 0.84 2.21 1.26 1.09 1.05 0.81 2.38 2.31 2.59 4.76 4.48 2.73 ...
         2.63 2.56 5.01 2.52 2.14 5.04 4.62 7.14]';
x2 = [2.12 2.07 2.97 2.45 3.18 2.07 2.41 4.97 4.76 6.58 6.44 3.36 5.04 ...
         3.60 6.47 7.31 2.77 4.48 6.30 4.73 4.66]';
```

Vi kan plotta y både mot x_1 och x_2.



Figur 2.5: Till vänster: y plottad mot x_1. Till höger: y plottad mot x_2.

```
figure
scatter(x1,y,'*')
xlabel('x_1'), ylabel('y')
figure
scatter(x2,y,'*')
xlabel('x_2'), ylabel('y')
```

Vi ser i Figur 2.5 att sambandet mellan y och x_1 ser ganska linjärt ut och vi har också en tendens till linjärt samband mellan y och x_2. Det skulle därför kunna vara intressant att beskriva priset y_j på hamburgare nummer j som en observation av en stokastisk variabel

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \varepsilon_j,$$

för $j = 1, \ldots, n$, där $\beta_0, \beta_1, \beta_2$ är okända parametrar, x_{j1}, x_{j2} är fixa tal och ε_j är oberoende stokastiska variabler med $\mathrm{E}(\varepsilon_j) = 0$ och $\mathrm{var}(\varepsilon_j) = \sigma^2$.

Vi skattar de okända parametrarna med MATLAB.

```
>> regr = regstats(y,[x1 x2],'linear','all');
>> betahat = regr.tstat.beta

betahat =

    0.1652
    0.4939
    0.0789
```

```
>> s = sqrt(regr.mse)

s =

    0.2015

>> dfe = regr.fstat.dfe

dfe =

    18
```

Vi ser att $\widehat{\boldsymbol{\beta}} = \begin{pmatrix}0.1652 & 0.4939 & 0.0789\end{pmatrix}'$ *och* $s = 0.2015$ *med 18 frihetsgrader.*

Vi ska nu konstruera ett 95% konfidensintervall för $\mathrm{E}(Y_0) = \beta_0 + 3.6\beta_1 + 4.2\beta_2 = \mu_0$ *och ett 95% prediktionsintervall för* $Y_0 = \beta_0 + 3.6\beta_1 + 4.2\beta_2 + \varepsilon_0$*. Låt*

$$\boldsymbol{u} = \begin{pmatrix}1 & 3.6 & 4.2\end{pmatrix}'$$

och konfidensintervallet ges av

$$I_{\mu_0} = \left(\hat{\mu}_0 \mp t_{0.975}(18)\sqrt{\boldsymbol{u}'\widehat{\boldsymbol{C}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{u}}\right),$$

där $\hat{\mu}_0 = \boldsymbol{u}'\widehat{\boldsymbol{\beta}} = 2.2745$ *och* $t_{0.975}(18) = 2.10$*. Vi behöver nu beräkna* $\widehat{\boldsymbol{C}}_{\widehat{\boldsymbol{\beta}}}$ *eller* $(\boldsymbol{X}'\boldsymbol{X})^{-1}$*.*

```
>> u = [1 3.6 4.2]';
>> Cbetahat = regr.covb   % Kovariansmatris för betahat

Cbetahat =

   0.014611346073971  -0.000234624773822  -0.002842681162981
  -0.000234624773822   0.001125358737283  -0.000699661849842
  -0.002842681162981  -0.000699661849842   0.001142226364443

>> t = tinv(0.975,dfe);
>> I_EY0 = [u'*betahat-t*sqrt(u'*Cbetahat*u), u'*betahat+t*sqrt(u'*Cbetahat*u)]

I_EY0 =

   2.1670   2.3820
```

Alltså, konfidensintervallet blir

$$I_{\mathrm{E}(Y_0)} = I_{\beta_0+3.6\beta_1+4.2\beta_2} = (2.17\ ;\ 2.38)$$

Vi vill också beräkna ett 95% prediktionsintervall för $Y_0 = \beta_0 + 3.6\beta_1 + 4.2\beta_2 + \varepsilon_0$*. Prediktionsintervallet ges av*

$$I_{Y_0} = \left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}} \mp t_{0.975}(18)\sqrt{s^2 + \boldsymbol{u}'\widehat{\boldsymbol{C}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{u}}\right).$$

```
>> I_Y0 = [u'*betahat-t*sqrt(s^2+u'*Cbetahat*u), ...
                          u'*betahat+t*sqrt(s^2+u'*Cbetahat*u)]

I_Y0 =

   1.8377   2.3625
```

Prediktionsintervallet blir

$$I_{Y_0} = (1.84\ ;\ 2.71)\,.$$

■

2.5 Val av regressionsmodell

När man studerar en responsvariabel Y, väljer man ofta mellan flera alternativa modeller genom att man kan göra olika val av förklaringsvariabler. Ambitionen är att förklara så stor del av variationerna i y-värdena som möjligt genom att utnyttja relevanta förklaringsvariabler.

2.5.1 Residualanalys

När man har valt en modell och skattat alla de ingående okända parametrarna så behöver man också kontrollera om förutsättningarna för regressionsmodellen är uppfyllda. Det finns många olika sätt att göra det på men enklast kan vara att börja med en *residualanalys*. Residualerna är de skattade felen, det vill säga observerade värden på ε_i. Antag att vi har modell given i (2.1). De skattade väntevärdena $\hat{\mu}_j$ kallas ibland också $\hat{y}_j$, det vill säga

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_k x_{jk}.$$

Residualerna definieras som

$$e_i = y_j - \hat{y}_j.$$

Om vi vill arbeta med vektorer istället så har vi modellen $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma \boldsymbol{I}_n)$ och residualerna ges av

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y}.$$

Enligt vårt modell antagande bör residualerna

1. ha konstant varians, oberoende av förklaringsvariablerna $x_1, \ldots, x_p$,
2. vara oberoende av varandra,
3. vara normalfördelade.

1. och 2. är viktiga för regressionsmodellen medan 3. är viktig för den fortsatta analysen när man ska göra inferens så som konfidensintervall och olika test. Alla våra t- och F-test bygger på normalfördelningsantagandet.

Enklast är att göra residualanalysen visuellt, det vill säga studera olika plottar. Det kan vara intressant att studera plottar av $e_1, \ldots, e_n$

1. på en tallinje (histogram),

2. i tidsföljd,

3. mot de skattade väntevärdena $\hat{\mu}_j$ (eller mot y_j),

4. mot var och en av förklaringsvariablerna x_j, $j = 1, \ldots, p$.

Exempel 13. *Vi har mätt hårdheten y för tolv stålplåtar för olika kombinationer av kopparinnehåll* x_1 *(enhet: %) och härdningstemperatur* x_2 *(enhet:* $100°F$*).*

```
y = [78.9 65.1 55.2 56.4 80.9 69.7 57.4 55.4 85.3 71.8 60.7 58.9]';
x1 = [0.02 0.02 0.02 0.02 0.1 0.1 0.1 0.1 0.18 0.18 0.18 0.18]';
x2 = [10 11 12 13 10 11 12 13 10 11 12 13]';
```

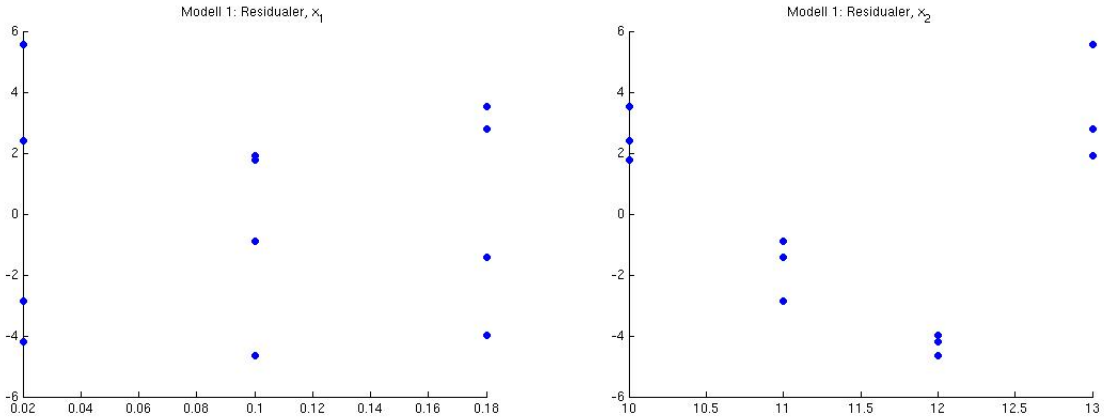
Med hjälp av MATLAB har en analys genomförts enligt

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad \textbf{(Modell 1)}.$$

```
% Modell 1
regr = regstats(y,[x1 x2],'linear','all');
betahat = regr.tstat.beta
fstat = regr.fstat

SSe1 = fstat.sse
dfe1 = fstat.dfe

R2 = regr.rsquare
e = regr.r;

figure; scatter(x1,e,'filled'); title('Modell 1: Residualer, x_1')
figure; scatter(x2,e,'filled'); title('Modell 1: Residualer, x_2')
```

vilket ger residualplottar i Figur 2.6. Vi ser att sambandet mellan y och x_2 *inte verkar linjärt. Som det verkar finns det en kvadratisk term i residualerna. Vi utökar därför modellen genom att ta med även* x_2^2 *som förklaringsvariabel*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon \qquad \textbf{(Modell 2)}.$$



Figur 2.6: Residualplottar (Modell 1) – e_j plåtade mot x_{j1} till vänster och mot x_{j2} till höger.

```
---------- Modell 1 ----------

betahat =

  161.3365
   32.9687
   -8.5500

fstat =

     sse: 129.3404
     dfe: 9
     dfr: 2
     ssr: 1.1522e+03
       f: 40.0868
    pval: 3.2963e-05

SSe1 =

  129.3404

dfe1 =

     9

R2 =

    0.8991
```

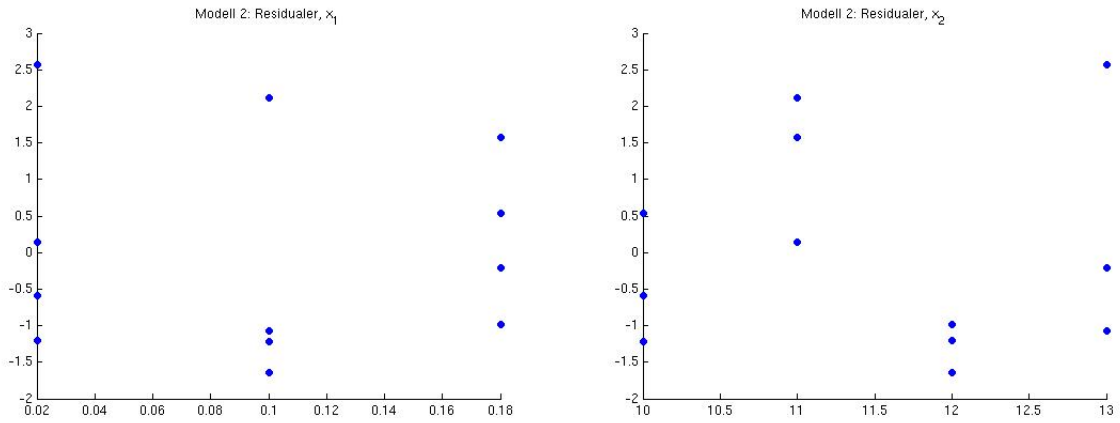
```
disp('---------- Modell 2 ----------')
x22 = x2.*x2;

regr2 = regstats(y,[x1 x2 x22],'linear','all');
betahat2 = regr2.tstat.beta

fstat2 = regr2.fstat
SSe2 = fstat2.sse
dfe2 = fstat2.dfe

R2 = regr2.rsquare
e2 = regr2.r;
figure; scatter(x1,e2,'filled'); title('Modell 2: Residualer, x_1')
figure; scatter(x2,e2,'filled'); title('Modell 2: Residualer, x_2')
```

vilket ger

```
---------- Modell 2 ----------

betahat2 =

  553.2448
   32.9687
  -77.3583
    2.9917

fstat2 =

     sse: 21.9396
     dfe: 8
     dfr: 3
     ssr: 1.2596e+03
       f: 153.0980
    pval: 2.0994e-07

SSe2 =

   21.9396

dfe2 =

     8

R2 =

    0.9829
```

Residualplottarna, Figur 2.7, är något bättre, men vi får vara observanta på att korrelationen mellan x_2 och x_2^2 är hög

```
>> corr(x2,x22)

ans =

    0.9992
```



Figur 2.7: Residualplottar (Modell 2) – e_j plåtade mot x_{j1} till vänster och mot x_{j2} till höger.

■

Det är värt att nämna att residualerna $e_1, \ldots, e_n$ approximerar $\varepsilon_1 \ldots, \varepsilon_n$ som ska vara oberoende. Residualerna $e_1, \ldots, e_n$ är däremot beroende. Detta beroende har dock inte så stor betydelse om kvoten $(n-k-1)/n$ är nära 1. Det beroendemönster som framstår i residualplottarna förklaras i allmänhet inte av beroendet mellan residualerna utan av att modellantagandet är felaktigt.

2.5.2 Jämförelse av två modeller

Modeller jämförs med hjälp av residualkvadratsummor respektive residualmedelkvadratsummor (= σ^2-skattningar). Det finns andra kriterier också men ett litet värde på σ antyder att ε-variabeln i modellen är ganska försumbar. En extra förklaringsvariabel ger **alltid** en minskning av residualkvadratsumman. Vi behöver kunna bedöma när denna minskning är signifikant.

Vi antar att vi vill jämföra

$$\textbf{Modell 1:} \quad Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

och

$$\textbf{Modell 2:} \quad Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k+p} x_{k+p} + \varepsilon$$

alltså vi vill eventuellt utvidga Modell 1 med p nya förklaringsvariabler. Man får i allmänhet nya β-koefficienter och en ny ε-variabel, men vi behåller samma notation för enkelhets skull.

Vi vill pröva

$$H_0: \ \beta_{k+1} = \ldots = \beta_{k+p} = 0 \text{ (nya förklaringsvariablerna meningslösa.)}$$

mot

$$H_1 : \text{minst en av } \beta_{k+1}, \ldots, \beta_{k+p} \neq 0.$$

Som teststorhet väljer vi

$$\boxed{W = \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)},}$$

där $SS_E^{(1)}$ är residualkvadratsumman för Modell 1 och $SS_E^{(2)}$ för Modell 2, $p =$ antalet nya förklaringsvariabler i Modell 2 och $n-k-p-1 = n-(k+p)-1$ är frihetsgraderna för $SS_E^{(2)}$. Eftersom en extra förklaringsvariabel alltid ger en minskning av residualkvadratsumman, det vill säga $SS_E^{(2)} < SS_E^{(1)}$, så vill vi förkasta H_0 om $W > c$. Vi har också att det för den stokastiska variabeln gäller att

$$W \sim F(p, n-k-p-1), \quad \text{om } H_0 \text{ är sann.}$$

Exempel 14. *Vi vill nu undersöka om Modell 2 är signifikant bättre än Modell 1 genom att göra ett F-test på nivån 0.001. Vi beräknar teststorheten genom*

```
>> w = ((SSe1 - SSe2)/1) / (SSe2/dfe2)

w =

   39.1624
```

och en kritisk gräns

```
>> c = finv(1-0.001,1,8)

c =

   25.4148
```

Eftersom $w = 39.16 > 25.41 = F_{0.999}(1,8) = c$ *förkastar vi* H_0*, om att* $\beta_3 = 0$*, det vill säga* x_2^2 *verkar göra nytta som förklaringsvariabel. Modell 2 verkar beskriva datamaterialet bättre än Modell 1.*

■

2.5.3 Olika metoder för stegvis regression

En vanlig situation är att man har observerade värden på en responsvariabel Y och ett antal förklaringsvariabler $x_1, \ldots, x_p$, men man vet inte vilka av dessa förklaringsvariabler som är relevanta och intressanta att ta med i modellen. Man vill använda tillräckligt många förklaringsvariabler för att förklara en stor del av variationen hör responsvariabeln, men man vill inte inkludera onödiga förklaringsvariabler och inte orimligt många. Frågan som uppstår då är: *Hur hittar man den 'bästa' modellen?* Svaret är tyvärr inte entydigt och det finns flera olika metoder för hur man ska välja uppsättning av förklaringsvariabler och metoderna ger inte alltid samma svar. Nedan följer en kort presentation av de enklaste metoderna.

Innan man sätter igång en metod bör man alltid omsorgsfullt överväga vilka förklaringsvariabler som är lämpliga att utnyttja.

Regression med alla delmängder

Om man har p förklaringsvariabler, så finns det 2^p olika modeller att välja mellan. Man kan genomföra analyser för alla dessa modellerna och till exempel välja den modell som ger minst variansskattning. Man kan även välja andra kriterier, se t.ex. [1].

Framåtvalsprincipen

Då man tillämpar stegvis regression börjar man ofta med en modell utan förklaringsvariabler och tar sedan in en förklaringsvariabel i sänder i modellen. För att välja en regressionsmodell enligt *framåtvalsprincipen*, väljer man först den förklaringsvariabel som har störst korrelation med y för att sedan lägga till en variabel i taget och man väljer den som ger minst residualkvadratsumma SS_E.

Om y_j är observation av $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$, för $j = 1, \ldots, n$, så kan man visa att

$$SS_R = r^2 \sum_{i=1}^{n} (y_j - \bar{y})^2,$$

$$SS_E = SS_{TOT} - SS_R = (1 - r^2) \sum_{i=1}^{n} (y_j - \bar{y})^2,$$

där r är den empiriska korrelationen för $(x_1, y_1), \ldots, (x_n, y_n)$. Den förklaringsvariabel som är starkast korrelerad med y (har högst värde på $|r|$) ger minst residualkvadratsumma vid analyser med **en** förklaringsvariabel. Observera att den näst bästa förklaringsvariabeln hittar man **inte** via korrelationsmatrisen utan vi måste fortsätta enligt följande.

1. Sök först upp den x-variabel som är **starkast korrelerad** med Y. Gör sedan en regressionsanalys med denna bästa ensamma förklaringsvariabel.

   Anteckna modellen, residualkvadratsumman (SS_E) och dess frihetsgrad. Undersök om "sambandet är signifikant", det vil säga avgör med ett test om β-koefficienten framför x-variabeln är skild från noll (vilket görs enklast med t-test). Om β-koefficienten är signifikant skild från noll är det klart att denna första förklaringsvariabel ska ingå i modellen.

2. Nästa steg är då att kombinera **var och en av de övriga** förklaringsvariablerna **med den först valda** och genomföra regressionsanalyser för att hitta det par, som bäst förklarar variationen hos Y, det cvill säga ger minst residualkvadratsumma (SS_E). Du behöver alltså göra $p - 1$ olika regressionsanalyser om det finns p förklaringsvariabler. Anteckna modellen, residualkvadratsumman och dess frihetsgrad för varje analys. Undersök för det bästa paret om den nya förklaringsvariabeln gör signifikant nytta.

3. Om även den andra förklaringsvariabeln du hittat skall ingå i modellen, så är nästa steg att man **studerar alla modeller med tre förklarande variabler**, där de båda först valda ingår.

   Fortsätt enligt framåtvalsprincipen och plocka in flera förklaringsvariabler i modellen om de gör nytta. Proceduren slutar då β-koefficienten framför den senaste förklaringsvariabeln inte är signifikant skild från noll.

Observera att de test som vi utför bara formellt har signifikansnivån α, eftersom vi hela tiden med hjälp av den observerade datan väljer vilka hypoteser som vi prövar.

Stegvis regression

Den metod som brukar kallas för *Stegvis regression* (eller *Stepwise* på engelska) är i princip samma som framåtvalsprincipen men så snart en ny variabel läggs till i modellen så testar man också om någon eller några av de tidigare ska tas bort genom att göra t- eller F-test. Signifikansnivåerna för att ta in (α_1) eller ta bort (α_2) kan vara olika, enligt $\alpha_1 \leq \alpha_2$, men oftast har man att de är lika, $\alpha_1 = \alpha_2$.

Bakåteliminationsprincipen

Bakåteliminationsprincipen fungerar enligt följande:

1. Genomför en regressionsanalys enligt den modell med alla p förklaringsvaraibler.
2. Testa sedan var och en av förklaringsvariablerna om de gör nytta, det vill säga testa hypoteserna $H_{0i} : \beta_i = 0$. Den förklaringsvariabel som har sämst värde på t-teststorheten genomför man det faktiska testet på.
   (a) Om man kan förkasta hypotesen ovan, så är proceduren avslutad och man väljer den modellen med alla p förklaringsvariabler.
   (b) Om man inte kan förkasta hypotesen så tar man bort den aktuella förklaringsvariabeln och börjar om i 1. med $p - 1$ förklaringsvariabler.

Litteraturförteckning

[1] Draper, N. R., & Smith, H. (1981). Applied regression analysis. 2nd Edition. *John Willey and Sons, New York, New York.*