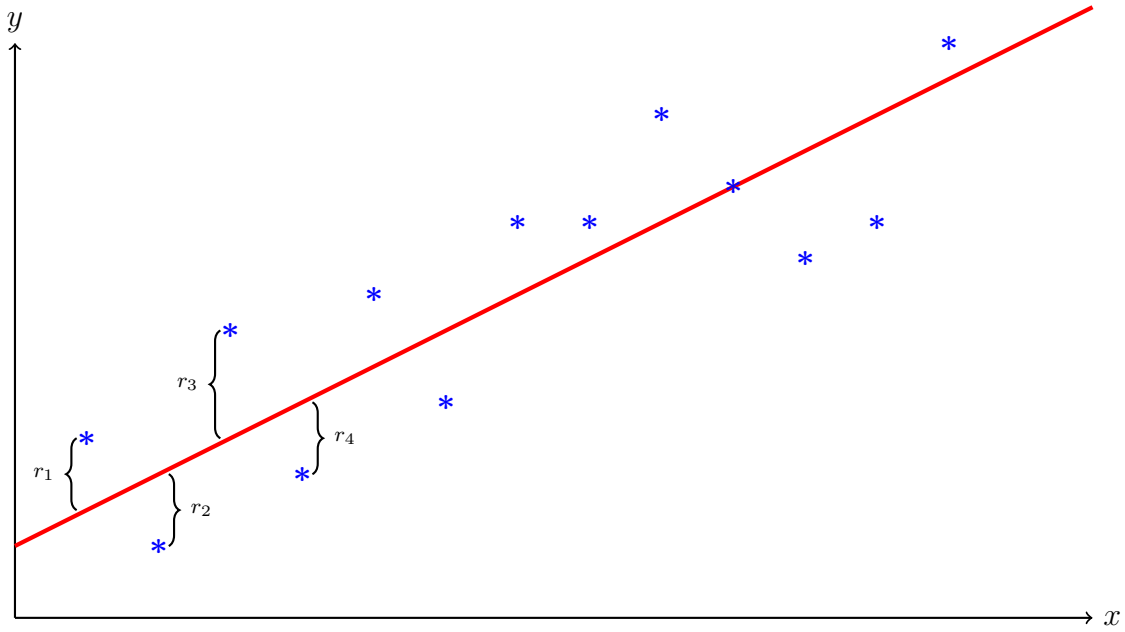
# Föreläsning 8: Linjär regression

Johan Thim (johan.thim@liu.se)

12 mars 2020

*"Your suffering will be legendary, even in hell."*
–Pinhead

Vi återgår nu till ett exempel vi stött på redan vid ett flertal tillfällen (föregående föreläsning och föreläsning 2 för att inte tala om tidigare kurser som linjär algebra), nämligen att anpassa en rät linje $y = \beta_0 + \beta_1 x$ efter mätdata $(x_i, y_i)$, $i = 1, 2, \ldots, n$. Grafiskt illustrerat enligt nedan.




Målsättningen är att – givet en mätserie – hitta den linje som approximerar denna serie på lämpligt sätt. Det resulterar i ett par naturliga funderingar.

(i) Hur hittar man en approximativ linje systematiskt?

(ii) Om man upprepar försöket, får man samma linje?

(iii) I vilken mening är linjen optimal? På vilket sätt mäter vi avvikelserna mellan linjen och mätserien?

Vi ska försöka svara på dessa frågor och för att göra det behöver vi ställa upp en modell.

## Enkel linjär regression

**Definition.** Vi kommer betrakta följande modell: givet $(x_j, y_j)$, $j = 1, 2, \ldots, n$, där vi betraktar $x_j$ som fixerade och $y_j$ som observationer av stokastiska variabler

$$Y_i = \beta_0 + \beta_1 x_j + \epsilon_j, \quad j = 1, 2, \ldots, n,$$

där $\epsilon_j \sim N(0, \sigma^2)$ antas oberoende (och likafördelade). Den räta linjen $y = \beta_0 + \beta_1 x$ kallas **regressionslinjen**.
Vi använder beteckningen $\mu_j = \beta_0 + \beta_1 x_j = E(Y_j)$.

Är det givet att denna modell är sann? Nej, det är inte självklart utan hänger på vilka förutsättningar datan kommer från. Däremot tenderar modellen att fungera bra i de flesta fall om vi har en rimlig mängd observationer.
Med notationen från figuren ser vi att

$$r_j = y_j - \mu_j$$

om vi tar hänsyn till tecknet (positivt tecken om $y_j$ ligger ovanför regressionslinjen). Ett mått på hur väl linjen approximerar mätserien ges av kvadratsumman

$$\sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - \mu_j)^2.$$

Vi skulle kunna minimera denna summa med avseende på $(\beta_0, \beta_1)$, vilket ger MK-skattningar för $\beta_0$ och $\beta_1$. Detta var det som hände i exemplet från föreläsning 2. Istället för att upprepa argumentet går vi över till den generella modellen för linjär regression.

## Linjär regression

**Definition.** Givet $(x_{j1}, x_{j2}, \ldots, x_{jk}, y_j)$, $j = 1, 2, \ldots, n$, för något positivt heltal $k$, där vi betraktar $x_{ji}$ som fixerade och $y_j$ som observationer av stokastiska variabler

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + \epsilon_j, \quad j = 1, 2, \ldots, n,$$

där $\epsilon_j \sim N(0, \sigma^2)$ antas oberoende (och likafördelade) och $\beta_0, \beta_1, \ldots, \beta_k$ är okända parameterar. Den räta linjen

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

kallas **regressionslinjen**.
Vi låter

$$\mu_j = E(Y_j) = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk}.$$

Dubbelindexeringen är lite jobbig att arbeta med, så låt oss gå över till matrisnotation:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 x_{11} & \cdots & \beta_k x_{1k} \\ \beta_0 & \beta_1 x_{21} & \cdots & \beta_k x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0 & \beta_1 x_{n1} & \cdots & \beta_k x_{nk} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Vi låter

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \text{samt } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

vilket leder till sambandet

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

Således gäller att

$$E(\boldsymbol{Y}) = X\boldsymbol{\beta} \quad \text{och} \quad C_{\boldsymbol{Y}} = \sigma^2 I_n,$$

där $I_n$ är den $n$-dimensionella enhetsmatrisen. Vi söker MK-skattningen $\widehat{\boldsymbol{\beta}}$ för $\boldsymbol{\beta}$, vilket vi kan erhålla genom att minimera den kvadratiska formen

$$Q(\beta_0, \ldots, \beta_k) = \sum_{j=1}^{n} (y_j - \mu_j)^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}),$$

där $\boldsymbol{y} = (y_1\ y_2\ \cdots\ y_n)^T$. En variant är att sätta igång och derivera, men lite mer elegant gäller följande sats.

## Normalekvationerna

**Sats.** Om $\det X^T X \neq 0$ så ges MK-skattningen av $\boldsymbol{\beta}$ enligt

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

**Bevis**. Vi kan skriva vektorn $\boldsymbol{Y}$ av uppmätta värden som

$$\boldsymbol{Y} = \beta_0 \boldsymbol{x_0} + \beta_1 \boldsymbol{x_1} + \cdots + \beta_k \boldsymbol{x_k} + \boldsymbol{\epsilon},$$

där $\boldsymbol{x_i}$, $i = 1, 2, \ldots, k$, är kolonn $i+1$ i matrisen $X$. Från linjär algebra vet vi att avståndet mellan observationen $\boldsymbol{y}$ och linjärkombinationer av vektorerna $\boldsymbol{x_j}$, dvs

$$\left| \boldsymbol{y} - (\widehat{\beta}_0 \boldsymbol{x_0} + \widehat{\beta}_1 \boldsymbol{x_1} + \cdots \widehat{\beta}_k \boldsymbol{x_k}) \right|,$$

blir minimalt precis då

$$X\widehat{\boldsymbol{\beta}} = \sum_{j=1}^{k} \widehat{\beta}_j \boldsymbol{x_j}$$

är projektionen av $\boldsymbol{y}$ på det linjära höljet span$\{\boldsymbol{x_0}, \boldsymbol{x_1}, \cdots, \boldsymbol{x_k}\}$. Således måste $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$ vara vinkelrät mot $\boldsymbol{x_j}$ för alla $j = 0, 1, \ldots, k$. Detta medför att

$$X^T(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = \boldsymbol{0} \quad \Leftrightarrow \quad X^T X\widehat{\boldsymbol{\beta}} = X^T \boldsymbol{y} \quad \Leftrightarrow \quad \widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y},$$

vilket är precis vad vi ville visa! □

När det gäller $\widehat{\boldsymbol{\beta}}$ och dess komponenter kommer vi använda samma beteckningar för observationer av punktskattningen och den stokastiska variabeln. Skulle vi vara konsekventa borde den stokastiska variabeln betecknas $\widehat{\boldsymbol{B}}$, men av tradition görs inte så. Var observant!

**Sats.** Med förutsättningarna ovan gäller att

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

**Bevis**. Det faktum att $\widehat{\boldsymbol{\beta}}$ är normalfördelad följer direkt från faktumet att elementen i $\widehat{\boldsymbol{\beta}}$ är linjärkombinationer av normalfördelade variabler. Återstår att visa väntevärde och kovarians:

$$E(\widehat{\boldsymbol{\beta}}) = (X^TX)^{-1}X^TE(\boldsymbol{Y}) = (X^TX)^{-1}X^TX\boldsymbol{\beta} = \boldsymbol{\beta}$$

och

$$\begin{aligned} C_{\widehat{\boldsymbol{\beta}}} &= (X^TX)^{-1}X^TC_{\boldsymbol{Y}}((X^TX)^{-1}X^T)^T = (X^TX)^{-1}X^T\sigma^2I_n((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^T\sigma^2I_n(X^T)^T((X^TX)^{-1})^T = \sigma^2(X^TX)^{-1}X^TX((X^TX)^T)^{-1} = \sigma^2(X^TX)^{-1}. \end{aligned}$$

Således blir fördelningen precis som beskriven i satsen. □

# 1 Variansanalys

Vi låter

$$\widehat{\mu_j} = \widehat{\beta}_0 + \widehat{\beta}_1x_{j1} + \widehat{\beta}_2x_{j2} + \cdots + \widehat{\beta}_kx_{jk}, \quad j = 1, 2, \ldots, n.$$

Ibland betecknas vektorn $\widehat{\boldsymbol{\mu}}$ som $\widehat{\boldsymbol{y}}$ eftersom det i någon mening är en skattningen av $\boldsymbol{y}$, men det blir lite olyckligt för det är inte $\boldsymbol{y}$ vi skattar. Vi ska nu studera hur bra den skattning vi tagit fram är.

**Regressionsanalysens kvadratsummor**

**Definition.** Vi definierar tre kvadratsummor som beskriver variationen hos $y$-värdena:

(i) Den **totala variationen** definieras enligt $\mathrm{SS_{TOT}} = \sum_{j=1}^{n}(y_j - \overline{y})^2$.

(ii) Variationen som förklaras av $x_1, x_2, \ldots, x_k$, definieras enligt $\mathrm{SS_R} = \sum_{j=1}^{n}(\widehat{\mu_j} - \overline{y})^2$.

(iii) Variationen som inte förklaras av regressionsmodellen är $\mathrm{SS_E} = \sum_{j=1}^{n}(y_j - \widehat{\mu_j})^2$.

Ibland används beteckningarna $Q_{\mathrm{TOT}}$ för $\mathrm{SS_{TOT}}$, $Q_{\mathrm{REGR}}$ för $\mathrm{SS_R}$ och $Q_{\mathrm{RES}}$ för $\mathrm{SS_E}$.
Hur hänger dessa summor ihop? Som tur är finns ett enkelt svar.

**Sats.** Den totala variationen $\mathrm{SS}_{\mathrm{TOT}}$ kan delas upp enligt

$$\mathrm{SS}_{\mathrm{TOT}} = \mathrm{SS}_{\mathrm{R}} + \mathrm{SS}_{\mathrm{E}}.$$

**Bevis**. Vi vill uttrycka $\mathrm{SS}_{\mathrm{TOT}}$ i termer av $\mathrm{SS}_{\mathrm{R}}$ och $\mathrm{SS}_{\mathrm{E}}$, så

$$\begin{aligned}
\mathrm{SS}_{\mathrm{TOT}} &= \sum_{j=1}^{n}(y_j - \overline{y})^2 = \sum_{j=1}^{n}(y_j - \widehat{\mu}_j + \widehat{\mu}_j - \overline{y})^2 \\
&= \sum_{j=1}^{n}(y_j - \widehat{\mu}_j)^2 + 2\sum_{j=1}^{n}(y_j - \widehat{\mu}_j)(\widehat{\mu}_j - \overline{y}) + \sum_{j=1}^{n}(\widehat{\mu}_j - \overline{y})^2 \\
&= \mathrm{SS}_{\mathrm{R}} + 2\sum_{j=1}^{n}(y_j - \widehat{\mu}_j)\widehat{\mu}_j - 2\overline{y}\sum_{j=1}^{n}(y_j - \widehat{\mu}_j) + \mathrm{SS}_{\mathrm{E}}.
\end{aligned}$$

Vi visar att de båda summorna i mitten summerar till noll. Eftersom $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$ gäller det att

$$\begin{aligned}
\sum_{j=1}^{n}(y_j - \widehat{\mu}_j)\widehat{\mu}_j &= \widehat{\boldsymbol{\mu}}^T(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = (X\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}}^T X^T(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) \\
&= \widehat{\boldsymbol{\beta}}^T(X^T\boldsymbol{y} - X^T X(X^T X)^{-1}X^T\boldsymbol{y}) = \widehat{\boldsymbol{\beta}}^T(X^T\boldsymbol{y} - X^T\boldsymbol{y}) = 0.
\end{aligned}$$

Med andra ord är $\boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ vinkelrät mot $\widehat{\boldsymbol{\mu}}$.

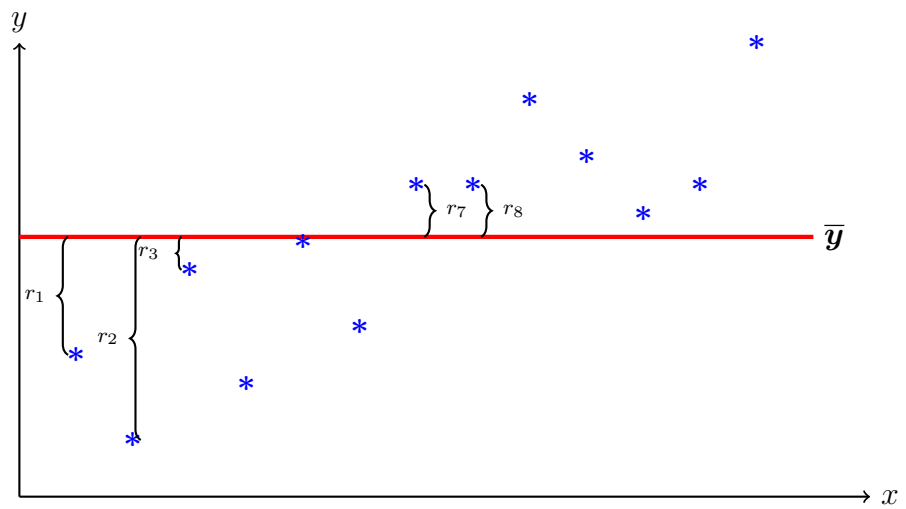
Vidare vet vi att $\widehat{\boldsymbol{\beta}}$ minimerar $Q(\boldsymbol{\beta}) = \sum_{j=1}^{n}(y_j - \mu_j)^2$, så

$$0 = \frac{\partial}{\partial_{\beta_0}} Q(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = -2\sum_{j=1}^{n}(y_j - \widehat{\mu}_j) \quad \Leftrightarrow \quad \sum_{j=1}^{n} y_j = \sum_{j=1}^{n}\widehat{\mu}_j,$$

vilket var precis det vi behövde. □

## 1.1 $\mathrm{SS}_{\mathrm{TOT}}$

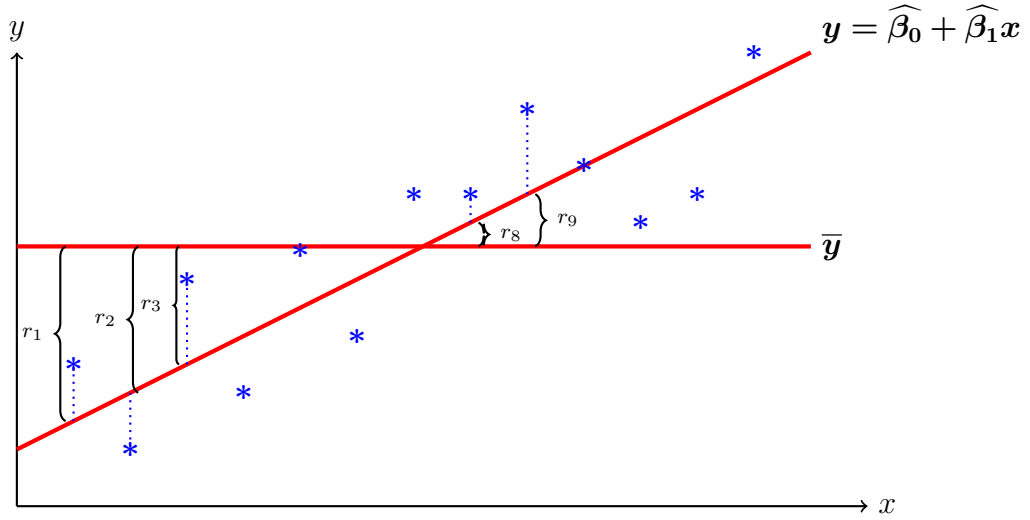
Storheten $\mathrm{SS}_{\mathrm{TOT}}$ mäter den totala variation av mätvärden jämfört med mätvärdenas medelvärde. I fallet då vi använder modellen $Y = \beta_0 + \epsilon$ ger således $\mathrm{SS}_{\mathrm{TOT}}$ hela felet i regressionen.



Vi ser att $\mathrm{SS_{TOT}} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - \overline{y})^2$.

## 1.2 $\mathrm{SS_R}$ och $\mathrm{SS_E}$

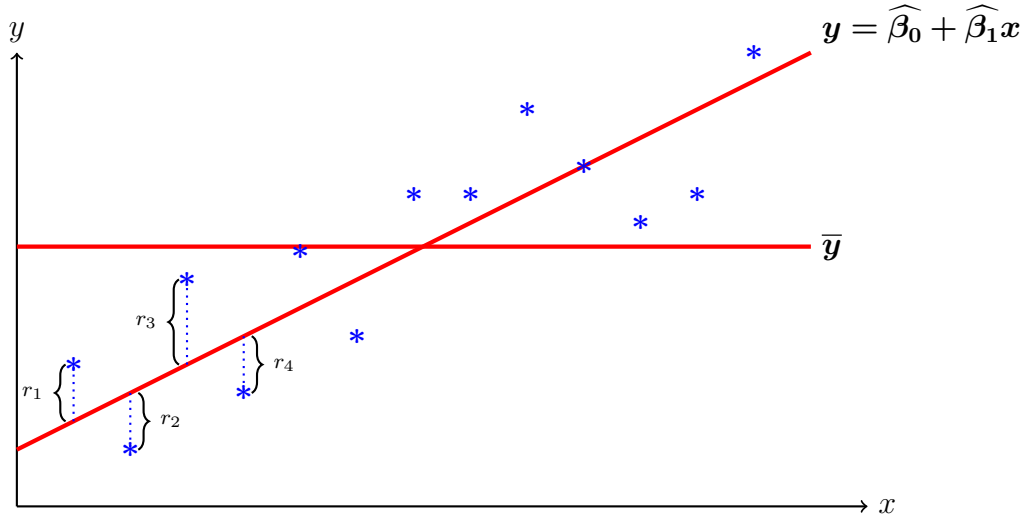
Om vi istället använder modellen $y = \beta_0 + \beta_1 x + \epsilon$ blir summorna $\mathrm{SS_E}$ och $\mathrm{SS_R}$ relevanta ($\mathrm{SS_{TOT}}$ ser fortfarande ut som ovan). Låt oss först illustrera $\mathrm{SS_R}$.




Vi ser att

$$\mathrm{SS_R} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 x_j - \overline{y})^2$$

och $\mathrm{SS_R}$ mäter alltså hur mycket den skattade regressionslinjen (i mätpunkterna $x_j$) skiljer sig från medelvärdet av $y$-värdena.
När det gäller $\mathrm{SS_E}$ är det istället skillnaden mellan den skattade regressionslinjen och mätvärdena vi betraktar.




Kvadratsumman av dessa avvikelser blir

$$\mathrm{SS_E} = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} (y_j - (\widehat{\beta}_0 + \widehat{\beta}_1 x_j))^2.$$

# 2 Projektionsmatriser

Eftersom vi arbetar med matriser och MK-lösningen i princip är projektionen på ett underrum av $\mathbf{R}^n$ så är följande resultat inte allt för förvånande.

**Hatt-matrisen**

**Definition.** Vi definierar $H = X(X^TX)^{-1}X^T$. Vi definierar även $J$ som en matris vars samtliga element är 1. Vi skriver $J_{nm}$ om vi vill markera dimensionen.

Varför kallar vi $H$ för hatt-matrisen? Ganska enkelt:

$$H\boldsymbol{y} = X(X^TX)^{-1}X^T\boldsymbol{y} = X\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{y}}.$$

Avbildningen sätter alltså hatten på!

**Sats.** Matriserna $H$, $I-H$ och $H-\frac{1}{n}J$ är projektionsmatriser $P$ som uppfyller $P^2 = P^T = P$.

**Bevis**. Direkt från definitionen av $H$ blir

$$H^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}X^T = H$$

och

$$H^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-T}X^T = X((X^TX)^T)^{-1}X^T = H.$$

Därav följer det att $(I-H)^2 = I^2 - 2H + H^2 = I - H$ och att $(I-H)^T = I^T - H^T = I - H$. För den sista operatorn skriver vi

$$\left(H - \frac{1}{n}J\right)^2 = H^2 - \frac{1}{n}HJ - \frac{1}{n}JH + \frac{1}{n^2}J^2 = H - \frac{1}{n}HJ - \frac{1}{n}JH + \frac{1}{n}J$$

ty $J^2$ är matrisen med samtliga element lika med $n$. Vidare gäller att $J$ kommuterar med alla kvadratiska matriser, så $JH = HJ$. Här kan vi se att $H\mathbf{1} = \mathbf{1}$ (där $\mathbf{1}$ är en vektor med samtliga element lika med 1) eftersom lösningen till regressionsproblemet om $\boldsymbol{y}$ är konstant är just den konstanten (lösningen är exakt). Alltså blir

$$\left(H - \frac{1}{n}J\right)^2 = H - \frac{1}{n}J.$$

Givetvis gäller även att $\left(H - \frac{1}{n}J\right)^T = H - \frac{1}{n}J.$ □

En följdsats av detta är att vi kan skriva kvadratsummorna som kvadratiska former.

**Sats.** $\mathrm{SS_{TOT}} = \boldsymbol{y}^T\left(I - \frac{1}{n}J\right)\boldsymbol{y}$, $\mathrm{SS_R} = \boldsymbol{y}^T\left(H - \frac{1}{n}J\right)\boldsymbol{y}$, samt $\mathrm{SS_E} = \boldsymbol{y}^T(I-H)\,\boldsymbol{y}$.

# 3 Förklaringsgrad

När vi utför regressionsanalys vill vi ofta ha ett mått som preciserar hur bra modellen passar de uppmätta värdena. Ett sådant mått är förklaringsgraden.

**Definition. Förklaringsgraden** $R^2$ definieras som

$$R^2 = \frac{\mathrm{SS_R}}{\mathrm{SS_{TOT}}} = \frac{\mathrm{SS_{TOT}} - \mathrm{SS_E}}{\mathrm{SS_{TOT}}} = 1 - \frac{\mathrm{SS_E}}{\mathrm{SS_{TOT}}}.$$

Ibland använder man lite andra varianter av $R^2$ och en mycket vanlig är den så kallade **justerade förkaringsgraden**

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SS_E}/(n-k-1)}{\mathrm{SS_{TOT}}/(n-1)}.$$

Om inte $n$ är förhållandevis stor i jämförelse med $k$ (vilket är nödvändigt för att regressionen ska vara vettig) så blir den justerade graden sämre.

# 4 Regressionsanalysens huvudsats

Innan vi ger oss in på huvudsatsen visar vi en hjälpsats från linjär algebra.

**Sats.** Låt $\boldsymbol{x} = (x_1\ x_2\ \cdots\ x_n)$ vara sådan att

$$\boldsymbol{x}^T\boldsymbol{x} = \sum_{j=1}^{n} x_j^2 = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{x}^T B\boldsymbol{x}$$

för $A, B \in \mathbf{R}^{n\times n}$ positivt semi-definita och symmetriska med $\mathrm{rank}(A) = r$ och $\mathrm{rank}(B) = n-r$ för något heltal $r$ så att $0 < r < n$. Då finns en ON-matris $C$ så att med $\boldsymbol{x} = C\boldsymbol{y}$ gäller att

$$\boldsymbol{x}^T A\boldsymbol{x} = \sum_{j=1}^{r} y_j^2 \quad \text{och} \quad \boldsymbol{x}^T B\boldsymbol{x} = \sum_{j=r+1}^{n} y_j^2.$$

**Bevis**. Eftersom $A$ är positivt semi-definit har $A$ endast icke-negativa egenvärden som vi ordnar enligt $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Vi vet även att $\mathrm{rank}(A) = r$, så $\lambda_{r+1} = \cdots = \lambda_n = 0$. Vidare är $A$ diagonaliserbar eftersom $A$ är symmetrisk, så det finns en ON-matris $C$ så att

$$C^T AC = D = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \vdots & \ddots & \vdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Med $\boldsymbol{x} = C\boldsymbol{y}$ gäller då att

$$\boldsymbol{x}^T\boldsymbol{x} = (C\boldsymbol{y})^T(C\boldsymbol{y}) = \boldsymbol{y}^T C^T C\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y}$$

och

$$\boldsymbol{x}^T A\boldsymbol{x} = (C\boldsymbol{y})^T AC\boldsymbol{y} = \boldsymbol{y}^T C^T AC\boldsymbol{y} = \boldsymbol{y}^T D\boldsymbol{y} = \sum_{j=1}^{r} \lambda_j y_j^2.$$

Vi har även

$$\boldsymbol{x}^T B\boldsymbol{x} = (C\boldsymbol{y})^T BC\boldsymbol{y} = \boldsymbol{y}^T C^T BC\boldsymbol{y}$$

och då blir

$$\sum_{j=1}^{n} y_j^2 = \boldsymbol{y}^T\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{x}^T B\boldsymbol{x} = \sum_{j=1}^{r} \lambda_j y_j^2 + \boldsymbol{y}^T C^T BC\boldsymbol{y}$$

så

$$\boldsymbol{y}^T C^T BC\boldsymbol{y} = \sum_{j=1}^{r} (1-\lambda_j) y_j^2 + \sum_{j=r+1}^{n} y_j^2.$$

Det faktum att $\operatorname{rank}(B) = n - r$ visar att $\lambda_1 = \lambda_2 = \cdots \lambda_r = 1$ och vi erhåller identiteten

$$\boldsymbol{y}^T C^T BC\boldsymbol{y} = \sum_{j=r}^{n} y_j^2.$$

Alla beteckningar är nu ur vägen och vi är framme vid huvudresultatet.

## Regressionsanalysens huvudsats

**Sats.** Med förutsättningarna ovan gäller följande för $\mathrm{SS_E}$ och $\mathrm{SS_R}$ sedda som stokastiska variabler.

(i) $\dfrac{\mathrm{SS_E}}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n} (Y_j - \widehat{\mu}_j)^2 \sim \chi^2(n-k-1).$

(ii) Givet att $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ är $\dfrac{\mathrm{SS_R}}{\sigma^2} = \dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n} (\widehat{\mu}_j - \overline{Y})^2 \sim \chi^2(k).$

(iii) Både $\mathrm{SS_R}$ och $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$ är oberoende av $\mathrm{SS_E}$.

**Bevis**. Eftersom $H$ är en projektionsmatris har $H$ endast egenvärdena $\lambda = 0$ och $\lambda = 1$. Således gäller det finurliga att matrisens rang är lika med summan av egenvärdena, vilka kan beräknas genom att ta spåret[1] av matrisen:

$$\operatorname{rank}(H) = \operatorname{tr}(H) = \operatorname{tr}(X(X^T X)^{-1} X^T) = \operatorname{tr}(X^T X(X^T X)^{-1}) = \operatorname{tr}(I_{k+1}) = k+1,$$

Det följer sedan att

$$\operatorname{rank}(I - H) = n - k - 1.$$

[1] se sista (bonus)avsnittet för lite detaljer kring spår av matriser.

Eftersom $HX = X$ så gäller att

$$\boldsymbol{Y}^T(I-H)\boldsymbol{Y} = \boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon}.$$

Detta är användbart eftersom $E(\boldsymbol{\epsilon}) = \mathbf{0}$. Enligt föregående hjälpsats gäller att sambandet

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T H\boldsymbol{\epsilon}$$

medför att det finns en ON-matris $C$ så att $\boldsymbol{Z} = C\boldsymbol{\epsilon}$ reducerar likheten till

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sum_{j=1}^{n-k-1} Z_j^2 + \sum_{j=n-k}^{n} Z_j^2$$

där

$$\boldsymbol{\epsilon}^T(I-H)\boldsymbol{\epsilon} = \boldsymbol{Y}^T(I-H)\boldsymbol{Y} = \mathrm{SS_E} = \sum_{j=1}^{n-k-1} Z_j^2.$$

Eftersom komponenterna i $\boldsymbol{\epsilon}$ är oberoende och $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$ följer det att

$$E(\boldsymbol{Z}) = C\mathbf{0} = \mathbf{0} \quad \text{och} \quad C_{\boldsymbol{Z}} = C_{C\boldsymbol{\epsilon}} = \sigma^2 C^T I C = \sigma^2 I$$

så $\boldsymbol{Z} \sim N\left(\mathbf{0}, \sigma^2 I\right)$. Komponenterna i $\boldsymbol{Z}$ är alltså oberoende och $Z_j \sim N(0, \sigma^2)$. Detta medför att

$$\frac{\mathrm{SS_E}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{n-k-1} Z_j^2 \sim \chi^2(n-k-1).$$

Vi erhåller även att $\mathrm{SS_E}$ och $H\boldsymbol{\epsilon}$ är oberoende (de delar inga variabler $Z_j$). Detta medför att $\mathrm{SS_E}$ och $\widehat{\boldsymbol{\beta}}$ är oberoende. Det faktum att $\mathrm{SS_E}$ och $\mathrm{SS_R}$ är oberoende följer av att kovariansen

$$\begin{aligned} C\left(\left(H-\frac{1}{n}J\right)\boldsymbol{Y}, (I-H)\boldsymbol{Y}\right) &= \left(H-\frac{1}{n}J\right)C(\boldsymbol{Y},\boldsymbol{Y})(I-H)^T = \sigma^2\left(H-\frac{1}{n}J\right)(I-H) \\ &= \sigma^2\left(H-H^2-\frac{1}{n}J+\frac{1}{n}JH\right) = \sigma^2\left(-\frac{1}{n}J+\frac{1}{n}HJ\right) \\ &= \sigma^2\left(-\frac{1}{n}J+\frac{1}{n}J\right) = 0, \end{aligned}$$

så dessa vektorer är okorrelerade och normalfördelade, så oberoende. Det följer direkt att $\mathrm{SS_E}$ och $\mathrm{SS_R}$ är oberoende (som funktioner av oberoende variabler).
Om vi fokuserar på fördelningen för $\mathrm{SS_R}$ så kan vi på samma sätt som ovan utnyttja att $\left(H-\frac{1}{n}J\right)$ är en projektionsmatris, så

$$\operatorname{rank}\left(H-\frac{1}{n}J\right) = \operatorname{tr}\left(H-\frac{1}{n}J\right) = \operatorname{tr}(H) - \operatorname{tr}\left(\frac{1}{n}J\right) = k+1-1 = k.$$

Matrisen $J$ är synnerligen rangdefekt med $\operatorname{rank}(J) = 1$ (eftersom alla kolonner är lika spänner vi bara upp ett en-dimensionellt rum).
Nu stöter vi på lite problem. Vi ser att

$$\left(H-\frac{1}{n}J\right)\boldsymbol{Y} = \left(I-\frac{1}{n}J\right)X\boldsymbol{\beta} + \left(H-\frac{1}{n}J\right)\boldsymbol{\epsilon}$$

där den första termen *inte* försvinner såvida inte $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. Men under detta antagande har vi åter igen en situation där vi kan "byta ut" $\boldsymbol{Y}$ mot $\boldsymbol{\epsilon}$. Föregående hjälpsats visar – analogt med föregående argument – att

$$\boldsymbol{Y}^T \left( H - \frac{1}{n} J \right) \boldsymbol{Y} = \sum_{j=1}^{k} Z_j^2,$$

där $Z_j$ är oberoende och $Z_j \sim N(0, \sigma^2)$, vilket medför att $\frac{1}{\sigma^2} \mathrm{SS_R} \sim \chi^2(k)$, under förutsättningen att $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. □

Kommentar: utan villkoret $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ kan man fortfarande genomföra argumentet, men resultatet blir en icke-centrerad $\chi^2(k)$-fördelning (något som inte ingår i kursen).

# 5 Hypotestester och kofidensintervall

Vi har nu samlat på oss en ordentlig verktygslåda, så det kanske är dags att se hur vi använder de olika delarna.

## 5.1 Skattning av $\boldsymbol{\sigma^2}$

Variansen $\sigma^2$ skattar vi med

$$s^2 = \frac{\mathrm{SS_E}}{n - k - 1}.$$

Denna skattning är väntevärdesriktig:

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

ty $\frac{1}{\sigma^2} \mathrm{SS_E} \sim \chi^2(n-k-1)$.

## 5.2 Finns det något vettigt i modellen?

En rimlig fråga efter utförd regression är om modellen faktiskt säger något. Vi brukar testa denna hypotes enligt följande. Vi testar nollhypotesen

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

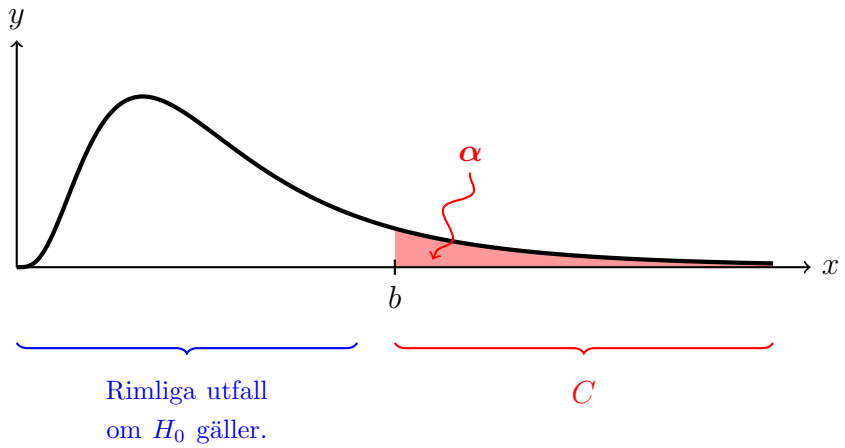
mot

$$H_1 : \text{minst något } \beta_i \neq 0.$$

En naturlig testvariabel ges av

$$v = \frac{\mathrm{SS_R}/k}{\mathrm{SS_E}/(n-k-1)}.$$

Huvudsatsen medför att $V \sim F(k, n-k-1)$ (om $H_0$ är sann) och vi förkastar $H_0$ om $v$ är stor. Hur stor avgörs av signifikansnivån och lite slagning i $F$-tabeller.




Vi hittar gränsen $b$ ur tabell (eller med `finv(1-alpha, k, n-k-1)` i MATLAB) så att

$$P(V > b) = \alpha.$$

## 5.3 Enskilda koefficienter

Eftersom $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}\right)$ introducerar vi beteckningen

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

Då är $\widehat{\beta}_i \sim N(\beta_i, \sigma^2 h_{ii})$. Om vi känner $\sigma^2$ kan vi använda att

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

för att testa hypotesen $H_0 : \beta_i = 0$ eller för att ställa upp konfidensintervall $I_{\beta_i}$.
Nu är det extremt sällan vi känner $\sigma^2$ exakt, men vi vet att $\widehat{\boldsymbol{\beta}}$ är oberoende av $\mathrm{SS_E}$ enligt huvudsatsen, så $\widehat{\beta}_i$ är oberoende av $\mathrm{SS_E}$. Vidare är $s^2 = \mathrm{SS_E}/(n-k-1)$ en skattning av $\sigma^2$ vi känner väl, så

$$\frac{(\widehat{\beta}_i - \beta_i)/(\sigma\sqrt{h_{ii}})}{\sqrt{((n-k-1)S^2/\sigma^2)/(n-k-1)}} = \frac{\widehat{\beta}_i - \beta_i}{S\sqrt{h_{ii}}} \sim t(n-k-1)$$

enligt Gossets sats.

## 5.4 Hypotestest: $H_0 : \beta_i = 0$

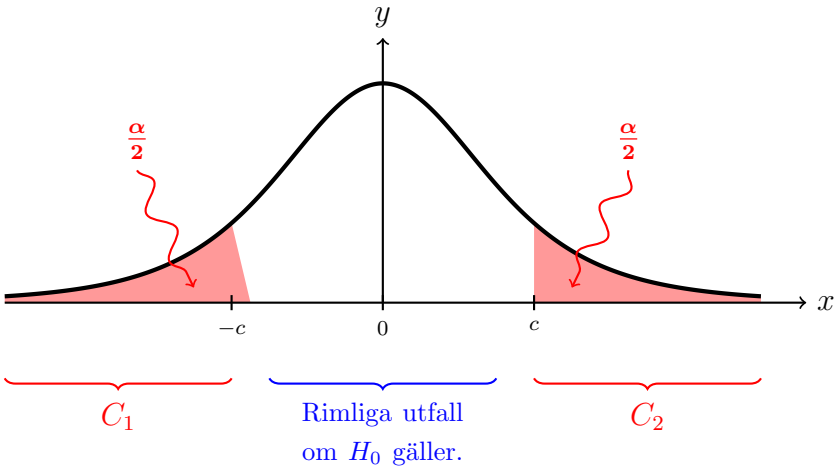
Vi kan även utföra hypotestester för en ensklid koefficient $\beta_i$ för att se om den förklaringsvariabeln tillför något signifikant (givetvis går det att ställa upp konfidensintervall också för mer kvantitativt innehåll).
Låt $H_0 : \beta_i = 0$ och $H_1 : \beta_i \neq 0$. Vi vet enligt ovan att

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

så det kritiska området finner vi enligt

$$C = \{t \in \mathbf{R} : |t| > c\}$$

för lämpligt tal $c > 0$ beroende på signifikansnivån och antalet frihetsgrader.




Gränsen hittar vi i tabell genom att leta reda på ett tal $c = F_T^{-1}(1 - \alpha/2)$ (sitter du med MATLAB kan du använda `c = -tinv(alpha/2)`).

# 6 Exempel: enkel linjär regression

Vi diskuterade enkel linjär regression tidigare i samband med MK-skattningar (föreläsning 2). Låt oss visa att vi får samma resultat med den metod vi nu tagit fram (och fördelningar för ingående storheter på ett enkelt sätt).
Vi låter $x_1, x_2, \ldots, x_n$ vara fixerade tal och $y_1, y_2, \ldots, y_n$ vara observationer från

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

där $\epsilon_i \sim N(0, \sigma^2)$ är oberoende. Alltså precis den modell vi använt tidigare. Syntesmatrisen $X$ ges av

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

så

$$X^T X = X = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \Rightarrow \quad (X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}$$

om $\det(X^T X) \neq 0$. Således blir

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} &= (X^T X)^{-1} X^T \boldsymbol{y} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix} \begin{pmatrix} n\overline{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (n\overline{x})^2} \begin{pmatrix} n\overline{y}\sum_{i=1}^n x_i^2 - n\overline{x}\sum_{i=1}^n x_i y_i \\ -n^2\overline{x}\overline{y} + n\sum_{i=1}^n x_i y_i \end{pmatrix}, \end{aligned}$$

vilket ger att

$$\widehat{\beta_1} = \frac{\sum_{j=1}^n x_j y_j - n\overline{x}\,\overline{y}}{\sum_{j=1}^n x_j^2 - n\overline{x}^2} = \frac{\sum_{j=1}^n y_j(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} = \frac{\sum_{j=1}^n (y_j - \overline{y})(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2} \quad \text{och} \quad \widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}.$$

Det följer nu att

$$\widehat{\beta_0} \sim N\left(\beta_0,\, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}\right) \quad \text{och} \quad \widehat{\beta_1} \sim N\left(\beta_1,\, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

Någonstans nu inser vi vilket kraftig syntaktiskt verktyg matriser och linjär algebra är..
Eftersom $\sigma$ är okänd skattar vi $\sigma^2$ med

$$s^2 = \frac{\mathrm{SS_E}}{n-2} = \frac{\sum_{j=1}^n (y_j - \widehat{\beta_0} - \widehat{\beta_1} x_j)^2}{n-2}$$

där $S^2 \sim \chi^2(n-2)$. Vi kan skriva om $\mathrm{SS_E}$ genom att utnyttja att

$$\begin{aligned}\mathrm{SS_R} &= \sum_{j=1}^n (\widehat{\mu_j} - \overline{y})^2 = \widehat{\beta_1}^2 \sum_{j=1}^n (x_j - \overline{x})^2 = \left(\frac{\sum_{j=1}^n (y_j - \overline{y})(x_j - \overline{x})}{\sum_{j=1}^n (x_j - \overline{x})^2}\right)^2 \sum_{j=1}^n (x_j - \overline{x})^2 \\ &= r^2 \sum_{j=1}^n (y_i - \overline{y})^2\end{aligned}$$

så

$$\mathrm{SS_E} = \mathrm{SS_{TOT}} - \mathrm{SS_R} = (1 - r^2) \sum_{j=1}^n (y_i - \overline{y})^2.$$

Vi kan här se att $r = 1$ ger perfekt matching så alla $(x_j, y_j)$ ligger på en rät linje.

# 7 Bonus: determinanter och spår

För en kvadratisk matris $A$ med dimension $n \times n$ definierar vi som bekant det **karakteristiska polynomet** enligt

$$p(\lambda) = \det(A - \lambda I).$$

Bekant från linjär algebra är att $p(\lambda)$ kan representeras enligt

$$\begin{aligned} p(\lambda) &= (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) \\ &= (-1)^n\big(\lambda^n - \lambda^{n-1}(\lambda_1 + \lambda_2 + \cdots + \lambda_n) + \cdots + (-1)^n \lambda_1\lambda_2\cdots\lambda_n\big), \end{aligned} \tag{1}$$

där $\lambda_i$, $i = 1, 2, \ldots, n$, är matrisens egenvärden.
**Spåret** för $A$ betecknar $\operatorname{tr} A$ och definieras som summan av diagonalelementen:

$$\operatorname{tr} A = \sum_{i=1}^n a_{ii}.$$

Spåret är linjär och uppfyller alltså att

$$\operatorname{tr}(cA) = c \operatorname{tr} A \quad \text{och} \quad \operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B.$$

Dessutom gäller att

$$\operatorname{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji}a_{ij} = \operatorname{tr}(BA)$$

om $A$ är $n \times m$ och $B$ är $m \times n$. Denna likhet leder direkt till att

$$\operatorname{tr}(PAP^{-1}) = \operatorname{tr}(P^{-1}(PA)) = \operatorname{tr} A,$$

vilket innebär att spåret (likt determinanten) är en invariant.
Vi vet även att $\det A = \lambda_1 \lambda_2 \cdots \lambda_n$, dvs determinanten ges av produkten av alla egenvärden. Finns något liknande för spåret? Svaret är vad man kanske skulle kunna gissa, nämligen att

$$\operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Enklaste beviset är att uttnyttja att alla matriser $A$ kan skrivas på Jordans normalform, dvs att det finns en basbytesmatris så att $J = PAP^{-1}$, där $J$ har $\lambda_1, \lambda_2, \ldots, \lambda_n$ på diagonalen och är nästan en diagonalmatris. Uttnyttjar vi att spåret är ivariant är vi klara. Kanske ligger detta argument lite utanför grundkursen i linjär algebra, så låt oss studera det karakteristiska polynomet lite närmare som alternativ. Genom att utveckla efter exempelvis rad 1 kan determinanten skrivas

$$p(\lambda) = (a_{11} - \lambda)\begin{vmatrix} a_{22} - \lambda & a_{23} & \cdots & a_{2n} \\ a_{32} & a_{33} - \lambda & \cdots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n2} & a_{n3} & \cdots & a_{nn} - \lambda \end{vmatrix} + p_1(t),$$

där $\operatorname{grad} p_1(\lambda) \leq n-2$ eftersom vi tagit bort en $\lambda$-term. Om vi utför samma operation på denna determinant ser vi att

$$p(\lambda) = (a_{11} - \lambda)(a_{22} - \lambda)\begin{vmatrix} a_{33} - \lambda & a_{34} & \cdots & a_{3n} \\ a_{43} & a_{44} - \lambda & \cdots & a_{4n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n3} & a_{n4} & \cdots & a_{nn} - \lambda \end{vmatrix} + p_2(t),$$

där $\operatorname{grad} p_2(\lambda) \leq n - 2$ med något (nytt) polynom $p_2(\lambda)$.
Upprepa proceduren $n$ gånger och vi erhållet att

$$\begin{aligned} p(\lambda) &= (a_{11} - \lambda)(a_{22} - \lambda)\cdots(a_{nn} - \lambda) + p_n(\lambda) \\ &= (-1)^n\lambda^n + (-1)^{n-1}\lambda^{n-1}(a_{11} + a_{22} + \cdots + a_{nn}) + p_{n+1}(\lambda) \\ &= (-1)^n\big(\lambda^n - \lambda^{n-1}\operatorname{tr} A\big) + p_{n+1}(\lambda), \end{aligned} \tag{2}$$

där $p_{n+1}(\lambda)$ är något polynom med grad högst $n - 2$. Om vi jämför (1) med (2) ser vi att

$$\operatorname{tr} A = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$