

Linear Curve of Regression of Y on X

$$\mu_{Y|x} = E[Y|x] = \beta_0 + \beta_1 x$$

Simple Linear Regression Model

Denote by $Y|x_i$ or Y_i the (random) observation at point x_i . Then

$$Y_i = \mu_{Y|x_i} + E_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \dots, n$$

where the error E_i has mean 0 and variance σ^2

Vi har alltså oberoende observationer $(x_1, y_1), \dots, (x_n, y_n)$ av (X, Y) och vi tänker oss dessa uppfyller

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

där ϵ_i är felet som görs i bestämningen av y_i

Minsta kvadrat-metoden

Så fort vi har skattningar b_0 och b_1 av β_0 och β_1 , så kan vi definiera residualerna

$$e_i = y_i - (b_0 + b_1 x_i)$$

Summan av residualernas kvadrater

$$\text{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

är ett mått på hur bra skattningarna b_0 och b_1 är.

Minsta kvadrat-skattningarna minimerar SSE.

Derivering m.a.p b_0 och b_1 ger

$$\frac{\partial \text{SSE}}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial \text{SSE}}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i$$

Sätter vi derivatorna till noll fås de s.k normalekvationerna

$$n b_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Dessa löses av minsta kvadrat-skattningarna.

Least-squares estimates for β_1 and β_0

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

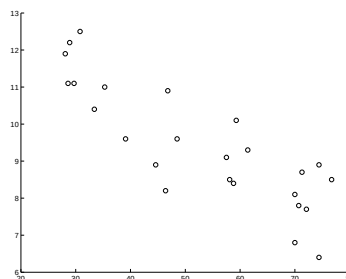
$$b_0 = \bar{y} - b_1 \bar{x}$$

Definiera dessutom

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

och lät $s = \sqrt{s^2}$.

Exempel 11.1.1



Data kan sammanfattas i

$$n = 25, \quad \sum x = 1314.90, \quad \sum y = 235.70$$

$$\sum x^2 = 76\,308.53, \quad \sum y^2 = 2286.07, \quad \sum xy = 11\,824.44$$

Observera också att

$$28.1 \leq x \leq 76.7$$

och

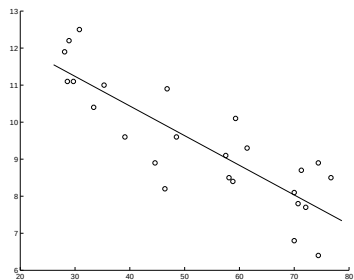
$$6.4 \leq y \leq 12.5$$

Insättning i formelerna för lutningen och "interceptet" ger

$$b_1 = -0.08 \quad \text{och} \quad b_0 = 13.64$$

Den estimerade regressionslinjen är således

$$\hat{\mu}_{Y|F} = \hat{y} = 13.64 - 0.08x$$



Med lite möda kan vi dessutom räkna ut att

$$s^2 = 0.7852 \quad \Rightarrow \quad s = 0.8861$$

Model assumptions: Simple linear regression

1. The random variables Y_i are independent and normally distributed
2. The mean of Y_i is $\beta_0 + \beta_1 x_i$
3. The variance of Y_i is σ^2

M.a.o

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$$

eller

$$Y | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$$

eller

$$Y | x \sim N(\beta_0 + \beta_1 x, \sigma)$$

Least-squares estimators for β_1 and β_0

$$\hat{\beta}_1 = B_1 = \frac{n \sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n Y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n x_i Y_i - n \bar{x} \bar{Y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$\hat{\beta}_0 = B_0 = \bar{Y} - B_1 \bar{x}$$

En väntevärdes riktig estimator av σ^2 är

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - B_0 - B_1 x_i)^2$$

Praktiska beteckningar

$$\begin{aligned} S_{xx} &= \sum_i (x_i - \bar{x})^2 \\ S_{yy} &= \sum_i (y_i - \bar{y})^2 \\ S_{xy} &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ \text{SSE} &= \sum_i (y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

Vi har redan sett att

$$S_{xx} = \sum x^2 - n \bar{x}^2 = \sum x^2 - \frac{1}{n} (\sum x)^2$$

Analogt visas att

$$\begin{aligned} S_{xy} &= \sum xy - n \bar{x} \bar{y} = \sum xy - \frac{1}{n} (\sum x) (\sum y) \\ S_{yy} &= \sum y^2 - n \bar{y}^2 = \sum y^2 - \frac{1}{n} (\sum y)^2 \end{aligned}$$

Således gäller

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

och det är inte särskilt svårt att visa att

$$\text{SSE} = S_{yy} - S_{xy}^2 / S_{xx}$$

$$B_1 \sim N\left(\beta_1, \sigma/\sqrt{S_{xx}}\right)$$

$$B_0 \sim N\left(\beta_0, \sigma\sqrt{\left(\frac{1}{n}\sum x^2\right)/S_{xx}}\right)$$

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$$

Dessutom gäller att

$$B_0, B_1 \text{ och } S^2 \text{ är oberoende}$$

Ur oberoendet och

$$\frac{B_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0, 1)$$

följer

$$\frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t(n-2)$$

Teststatistika vid test av $H_0 : \beta_1 = 0$ är således $\frac{B_1}{S/\sqrt{S_{xx}}}$

Exempel 11.1.1 (forts)

Vi beräknar

$$S_{xx} = \sum x^2 - \frac{1}{n}(\sum x)^2 = 7150.0496$$

$$S_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y) = -572.4372$$

$$S_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2 = 63.8904$$

samt

$$\text{SSE} = S_{yy} - S_{xy}^2/S_{xx} = 18.0607$$

och erhåller

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{18.0607}{23} = 0.7852 = 0.8861^2$$

Test av $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$

Observerat värde av teststatistikan är

$$\frac{b_1}{s/\sqrt{S_{xx}}} = -7.64$$

Antalet frihetsgrader är $n-2 = 23$ och i motsv t -tabell ser vi att $t_{0.005} = 2.807 < 7.64$.

P -värdet är således < 0.01 ($= 2 \cdot 0.005$ ty två-sidigt test).

Att $\beta_1 \neq 0$ är följaktligen statistiskt säkerställt.

Pearson correlation coefficient

$$\rho = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

Från teorin före bivariat N-fördelning hämtar vi

$$\mu_{Y|x} = \mu_Y + \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

och

$$\text{Var}[Y|x] = (1 - \rho^2)\sigma_y^2$$

Vi ser att

$$\beta_0 = \mu_Y - \rho \frac{\sigma_y}{\sigma_x} \mu_x, \quad \beta_1 = \rho \frac{\sigma_y}{\sigma_x}$$

$$\sigma^2 = (1 - \rho^2)\sigma_y^2$$

Således gäller

$$\rho = \beta_1 \frac{\sigma_x}{\sigma_y}$$

Estimator for ρ

$$\hat{\rho} = R = B_1 \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Notera att $\rho = 0 \Leftrightarrow \beta_1 = 0$ och att $\rho > 0 \Leftrightarrow \beta_1 > 0$

Således gäller att test av $H_0 : \rho = 0$ mot t.ex $H_1 : \rho \neq 0$ är ekvivalent med att testa $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$

Vi kan därför använda teststatistikan

$$\frac{B_1}{S/\sqrt{S_{xx}}} \sim t(n-2)$$

till att även testa $H_0 : \rho = 0$

Notera nu först att $B_1 = S_{XY}/S_{XX}$ och att

$$(n-2)S^2 = S_{YY} - S_{XY}^2/S_{XX} = S_{YY}(1 - R^2)$$

Således gäller att

$$\frac{B_1}{S/\sqrt{S_{xx}}} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

Alltså

Under $H_0 : \rho = 0$ är

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t(n-2)$$