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

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

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

Vi har alltså oberoende observationer $(x_1, y_1), \ldots, (x_n, y_n)$ 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

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

Sedan adderas residualernas kvadrater:

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

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

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

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

Least-squares estimates for β_1 and β_0

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

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

Ur 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 $28.1 \le x \le 76.7$ och $6.4 \le y \le 12.5$

Insättning ger $b_1 = -0.08006$ och $b_0 = 13.64$

Således är den estimerade regressionslinjen

$$\hat{\mu}_{Y|x} = \hat{y} = 13.64 - 0.08006x$$

T.ex gäller då $x = 50$ att $\hat{y} = 13.64 - 0.08006 \cdot 50 = 9.64$

Notera dessutom

$$x = 28.1 \;\Rightarrow\; \hat{y} = 11.39$$
$$x = 76.7 \;\Rightarrow\; \hat{y} = 7.50$$

Least-squares estimators for β_1 and β_0

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

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

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

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

Eller

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

Eller

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

Praktiska beteckningar

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

Motsv stokastiska variabler har analoga beteckningar

Notera exempelvis

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \sum_{i=1}^{n} x_i + n\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)$$

Det analoga uttrycket för S_{xx} har vi härlett i grundkursen och det finns såklart ett analogt uttryck för S_{yy}

Således gäller

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

Distribution of B_1

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

Distribution of B_0

$$B_0 \sim \mathrm{N}\left(\beta_0, \sigma \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}\right)$$

Estimator for σ^2

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

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

Notice

B_0, B_1 and S^2 are independent

Notera vidare

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

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

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

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

$$= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx}$$

Men

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

Således gäller

$$\text{SSE} = S_{yy} - b_1 S_{xy} = S_{yy} - S_{xy}^2 / S_{xx}$$

Forts av exempel 11.1.1

Vi beräknar

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

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

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

och

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

Vi ser att

$$s^2 = \frac{18.0607}{23} = 0.7852 = 0.8861^2$$

Regressionslinjens lutning $B_1 = \widehat{\beta}$

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

$$\Rightarrow \quad \frac{B_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathrm{N}(0,1) \quad \Rightarrow \quad \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t(n-2)$$

Forts av exempel 11.1.1

Observerat värde av B_1 är $b_1 = -0.08006$

Vi jämför med utfallet av $S/\sqrt{S_{xx}}$, som är $s/\sqrt{S_{xx}} = 0.0105$

Alltså, $b_1/(s/\sqrt{S_{xx}}) = 7.64$

J.f.r med $P(T_{23} > 2.81) = 0.005$

P-värdet är således betydligt mindre än 1% (två-sidigt test)

Att $\beta_1 \neq 0$ kan följaktligen anses i det närmaste säkerställt

Regressionslinjen $\hat{\mu}_{Y|x} = \widehat{Y} = B_0 + B_1 x$

$$\widehat{Y} \sim \mathrm{N}\left(\beta_0 + \beta_1 x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

$$\Rightarrow \quad \frac{\widehat{Y} - (\beta_0 + \beta_1 x)}{S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

Prediktionsfel

För givet värde på x är $Y|x \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma)$

Antag att vi vill prediktera en ny observation $Y|x$ med

$$\widehat{Y} = B_0 + B_1 x$$

Då blir prediktionsfelet

$$Y|x - (B_0 + B_1 x) \sim \mathrm{N}\left(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

Motsv prediktionsintervall är:

$$Y|x = b_0 + b_1 x \pm t_{\alpha/2, n-2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

Pearson correlation coefficient

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

Från teorin före bivariat N-fördelning hämtar vi

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

och

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

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

$$R = \hat{\rho} = B_1\sqrt{\frac{S_{XX}}{S_{YY}}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

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

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