I det vi kallar (enkel) linjär regression har vi

Data

Oberoende observationer $(x_1, y_1), \ldots, (x_n, y_n)$
som vi tänker oss uppfyller

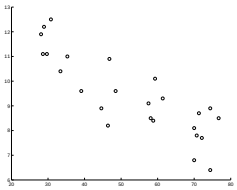
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

där ϵ_i har väntevärdet 0 och standardavvikelsen σ.

Parametrar

$$\beta_0, \quad \beta_1 \quad \text{och} \quad \sigma$$

Så här skulle det kunna se ut:



och vårt problem är att hitta den räta linje som bäst passar data.

Linear Curve of Regression of Y on X

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

$$\text{Var}[Y|x] = \sigma^2$$

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.

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

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

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

Normalekvationerna löses av minsta kvadrat-skattningarna.

Least-squares estimates for β_1 and β_0

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

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

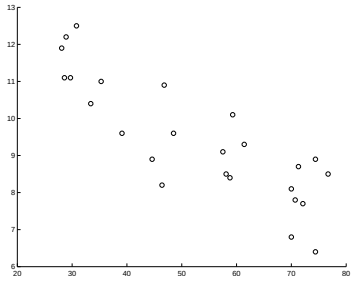
Definiera dessutom

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

och låt $s = \sqrt{s^2}$.

Exempel 11.1.1

x: relative humidity (%); y: solvent evaporation (% wt)



Data kan sammanfattas i

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

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

Observera också att

$$28.1 \leq x \leq 76.7$$

och

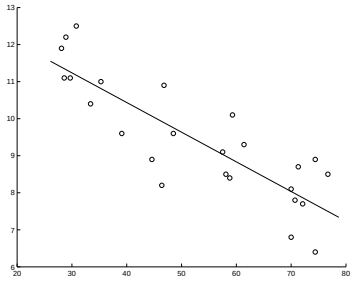
$$6.4 \leq y \leq 12.5$$

Insättning i formlerna för lutningen och "interceptet" ger

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

Den estimerade regressionslinjen är således

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



Med lite möda kan vi dessutom räkna ut att

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

T.ex kan vi nu prediktera avdunstningen då $x = 50$. Vi får

$$\hat{\mu}_{Y|x=50} = 13.64 - 0.08 \cdot 50 = 9.64$$

och vi vet dessutom att standardavvikelsen är ca ≈ 0.89.

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

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

eller

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

eller

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

Least-squares estimators for β_1 and β_0

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

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

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

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

Vi har redan sett att

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

Analogt visas att

$$S_{xy} = \sum xy - n\bar{x}\bar{y} = \sum xy - \tfrac{1}{n}\left(\sum x\right)\left(\sum y\right)$$
$$S_{yy} = \sum y^2 - n\bar{y}^2 = \sum y^2 - \tfrac{1}{n}\left(\sum y\right)^2$$

Således gäller

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

Dessutom gäller

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

Det är inte särskilt svårt att visa att

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

och vi inser att

$$s^2 = \frac{1}{n-2}\left(S_{yy} - S_{xy}^2/S_{xx}\right)$$

Distribution of B_1, B_0 and S^2

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

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

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

Dessutom gäller att

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

Ur oberoendet och

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

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 $\dfrac{B_1}{S/\sqrt{S_{xx}}}$.

Oftast testar man mot alternativet $H_1 : \beta_1 \neq 0$.

Nivå-α-testregel är förkasta då $\left|\dfrac{B_1}{S/\sqrt{S_{xx}}}\right| \geq t_{\alpha/2}(n-2)$.

Exempel 11.1.1 (forts)

Vi beräknar

$$S_{xx} = \sum x^2 - \tfrac{1}{n}\left(\sum x\right)^2 = 7150.0496$$
$$S_{xy} = \sum xy - \tfrac{1}{n}\left(\sum x\right)\left(\sum y\right) = -572.4372$$
$$S_{yy} = \sum y^2 - \tfrac{1}{n}\left(\sum y\right)^2 = 63.8904$$

samt

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

och erhåller

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

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

Observerat värde av testatistikan 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, vilket vi även hade kunnat konstatera genom att beräkna 99%-intervallet

$$\begin{aligned} \beta_1 &= -0.08 \pm 2.807 \cdot 0.8861/\sqrt{7150.0496} \\ &= -0.08 \pm 0.03 \end{aligned}$$

Nu ändrar vi modellen en aning och tänker oss att vi har oberoende observationer x_i, y_i av en bivariat normalfördelning med correlation ρ.

Pearson correlation coefficient

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

Från teorin för 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

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

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 $H_0 : \rho = 0$

Men notera 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

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

Alltså

Under $H_0 : \rho = 0$ är

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

Exempel 11.6.1

Man jämför en äldre manuell metod för att mäta koncentrationen kväve i vatten, x, med en ny automatisk metod, y, och vill visa att $\rho_{xy} > 0$.

Data sammanfattas av:

$$n = 10, \quad \sum x = 2405, \quad \sum y = 2503$$
$$\sum x^2 = 900\,775, \quad \sum y^2 = 919\,489, \quad \sum xy = 902\,475$$

Vi räknar ut att

$$S_{xx} = 322\,372.5, \quad S_{xy} = 300\,503.5, \quad S_{yy} = 292\,988.1$$

Och får skattningen

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{300\,503.5}{\sqrt{322\,372.5 \cdot 292\,988.1}} = 0.978$$

Observerat värde av teststatistikan är

$$\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} = \frac{0.978\sqrt{10-2}}{\sqrt{1-0.978^2}} = 13.26$$

som vi vet är $t(8)$-fördelad. Ur tabell fås $t_{0.005} = 5.041 < 13.26$.

Slutsats: P-värdet är betrydligt mindre än 0.005 och vi kan tryggt konstatera att $\rho_{xy} > 0$.