# Least squares

Data

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Quantities $X$ and $Y$

Relationship $Y = \beta_0 + \beta_1 X$ ?

Model relationship

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

$\beta_0, \beta_1$ - coefficients of the model (in $\mathbb{R}$)

$e_i$ - unobservable errors

Goal: find BEST $\beta_0$ and $\beta_1$ (estimates) labelled $\hat{\beta}_0$ and $\hat{\beta}_1$

$$e_i = y_i - \beta_0 - \beta_1 x_i$$

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

observable errors

$\sum_{i=1}^{n} (\hat{e}_i)^2$ is minimised.

Residual Sum of squares (RSS)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i]^2$$

Set $z_i = y_i - \beta_1 x_i$

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [z_i - \beta_0]^2 \geq \sum_{i=1}^{n} [z_i - \bar{z}]^2$$

Exercise

$$\Rightarrow \hat{\beta}_0 = \bar{z} = \bar{y} - \beta_1 \bar{x}$$

$$= \frac{1}{n} \sum_{i=1}^{n} [y_i - \beta_1 x_i]$$

then $RSS(\hat{\beta}_0, \beta_1) \leq RSS(\beta_0, \beta_1)$

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - \beta_1 x_i - \beta_0]^2$$

$$RSS(\hat{\beta}_0, \beta_1) = \sum_{i=1}^{n} [y_i - \beta_1 x_i - \hat{\beta}_0]^2$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

$$RSS(\hat{\beta}_0, \beta_1) = \sum_{i=1}^{n} [y_i - \beta_1 x_i - (\bar{y} - \beta_1 \bar{x})]^2$$

$$= \sum_{i=1}^{n} [(y_i - \bar{y}) - \beta_1 (x_i - \bar{x})]^2$$

$$= \sum_{i=1}^{n} \left[ (y_i - \bar{y})^2 - 2\beta_1 (x_i - \bar{x})(y_i - \bar{y}) + \beta_1^2 (x_i - \bar{x})^2 \right]$$

$$= \sum_{i=1}^{n} (y - \bar{y})^2 - 2\beta_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

$$RSS(\hat{\beta}_0, \beta_1) = S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx}$$

Complete the square

$$= S_{xx} \left( \beta_1^2 - 2\beta_1 \frac{S_{xy}}{S_{xx}} \right) + S_{yy}$$

$$= S_{xx} \left( \beta_1^2 - 2\beta_1 \frac{S_{xy}}{S_{xx}} + \left( \frac{S_{xy}}{S_{xx}} \right)^2 \right) + S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$= S_{xx} \left( \beta_1 - \frac{S_{xy}}{S_{xx}} \right)^2 + \left( S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \right)$$

$$\geq \left( S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \right)$$

minimised when $\beta_1 = \frac{S_{xy}}{S_{xx}}$

So: $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ minimise $RSS(\beta_0, \beta_1)$

---

## Simple linear regression

X, Y

Data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

We assume:

$x_i$ are fixed

$\beta_0, \beta_1$ are fixed, unknown parameters

$\epsilon_i$ are INDEPENDENT errors

$\epsilon_i \sim N(0, \sigma^2)$, $\sigma^2$ unknown.

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

LINEAR regression

functions $f, g$

$$f(Y) = \beta_0 + \beta_1 g(x_i) + \epsilon_i$$

also linear regression.

$$Y = \beta_0 + \beta_1 x_i^2 + \epsilon_i \quad \checkmark$$

$$\log Y = \beta_0 + \beta_1 \sqrt{x_i} + \epsilon_i \quad \checkmark$$

$$Y = \beta_0 + \beta_1 \exp(x_i) + \epsilon_i \quad \checkmark$$

$$Y = \beta_0 \exp(\beta_1 x_i) + \epsilon_i \quad X$$

---

Simple linear regression

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

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \; ; \; \sigma^2 \text{ unknown}$$

$$Y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$E[Y_i] = \beta_0 + \beta_1 x_i$$

$$Var(Y_i) = \sigma^2$$

Estimate parameters $\beta_0$ and $\beta_1$

$$f(y_i | \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta x_i)^2\right]$$

$$y = (y_1, y_2, \ldots, y_n)$$

$$f(y | \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i | \beta_0, \beta_1, \sigma^2)$$

$$L(\beta_0, \beta_1, \sigma^2 | y) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta x_i)^2\right)$$

Maximise likelihood

$$\log L(\beta_0, \beta_1, \sigma^2 | y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}(y_i - \beta_0 - \beta_1 x_i)^2$$

Assume $\sigma^2$ fixed for now

need to maximise: $\sum_{i=1}(y_i - \beta_0 - \beta_1 x_i)^2$

We have done this already!

$$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Next: maximise $\log L$ w.r.t. $\sigma^2$

$$\log L(\sigma^2 | x, y, \hat{\beta}_0, \hat{\beta}_1)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Let $\sigma^2 = w \Rightarrow -\frac{d}{dw} \Rightarrow \hat{\sigma}^2$

Fit model with $\hat{\beta}_0$ and $\hat{\beta}_1$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$

$\hat{\epsilon}_i$ - (observable) residuals