Regression - Part 2

Simple linear regression

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

Model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2, \ldots, n$

Assumptions

β_0, β_1 are fixed, unknown parameters

x_i fixed values

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

ϵ_i are independent

Can find MLEs of β_0, β_1

$$\boxed{\begin{aligned} \hat{\beta}_0 &= \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \end{aligned}}$$

Agree with least squares approach.

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

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

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

Key question: how do we know if our model is a "good"?

Plot $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ $\Leftarrow$ fitted (regression line)

But what about ϵ_i's?

$\hat{\beta}_0$ is an estimate of β_0

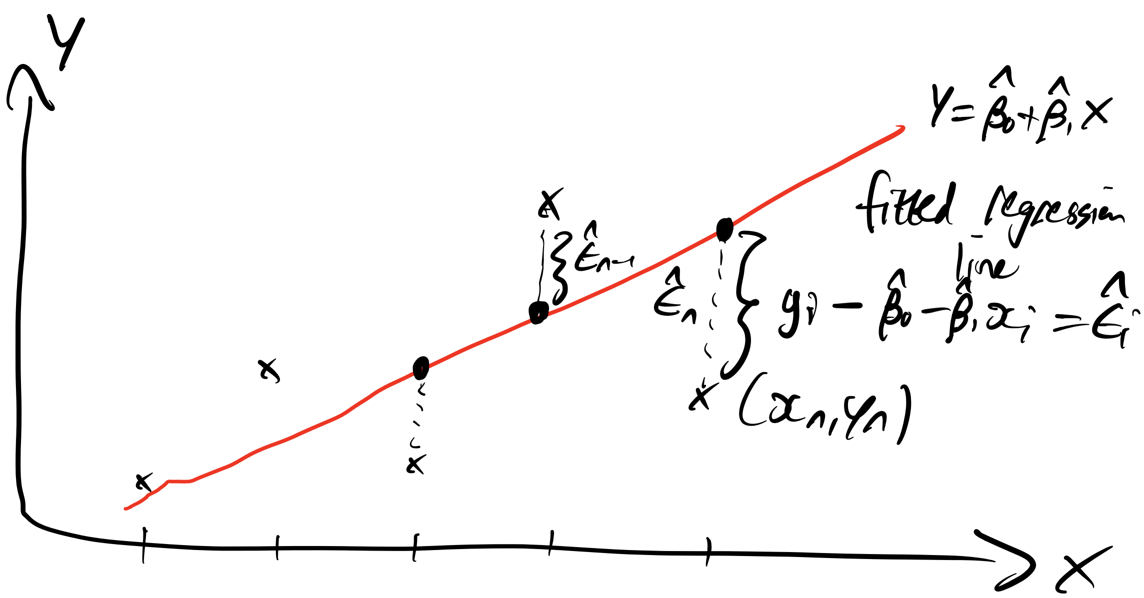
$\hat{\beta}_1$ is an estimate of β_1

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i \; ?$$

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

estimates of ϵ_i

We call these residuals



Option 1 $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$

Option 2 : $Y_i = \beta_0 + \beta_2 x_i^2 + \epsilon_i$

Option 3: $f(Y_i) = \beta_0 + \beta_1 g(x_i) + \epsilon$

for functions f and g

Residuals sum of squares

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad i = 1, 2, \ldots, n$$

$$\widehat{RSS}_{xy} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

(we minimised $RSS = \sum_{i=1}^{n} \epsilon_i^2$)

$$\widehat{RSS}_{xy} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

(Exercise 9.4.3)

$$\widehat{RSS}_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

<u>Definition 9.1</u>: The R-squared statistic R^2 is

$$R^2 = \frac{S_{yy} - \widehat{RSS}_{xy}}{S_{yy}}$$

$$R^2 = \frac{1}{S_{yy}}\left[S_{yy} - \left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}}\right)\right]$$

$$= \frac{1}{S_{yy}}\left[\frac{(S_{xy})^2}{S_{xx}}\right]$$

$$= \frac{(S_{xy})^2}{S_{xx}S_{yy}}$$

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

$$= r_{xy}^2$$

$\rho_{xy} \in [-1, 1]$ (Result)

$r_{xy} \in [-1, 1]$ (Cauchy-Schwartz)

$r_{xy}^2 \in [0, 1]$