Module 3:
Linear Models I

Recap of simple linear regression (SLR)

deterministic models: y is completely specified for a given value of x, no randomness or error allowed

probablistic models: incorporates randomness

Regression analysis uses probablistic models to investigate the relationship between two or more variables.

Setup

Consider data of the form

predictor variable
(explanatory variable, independent variable)

response variable
(dependent variable)

$$\underline{(x_1, y_1)}, (x_2, y_2), \ldots, (x_n, y_n).$$

The simple linear regression model is

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{E(Y_i)} + \underbrace{\epsilon_i}_{\text{random component}} \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

deterministic component

with

$$\underline{\epsilon_i \sim N(0, \sigma^2)}$$

independently for $i = 1, 2, \ldots, n$.

e.g. $Y_i = \beta_0 + x_i^{\beta_1} + \varepsilon_i$

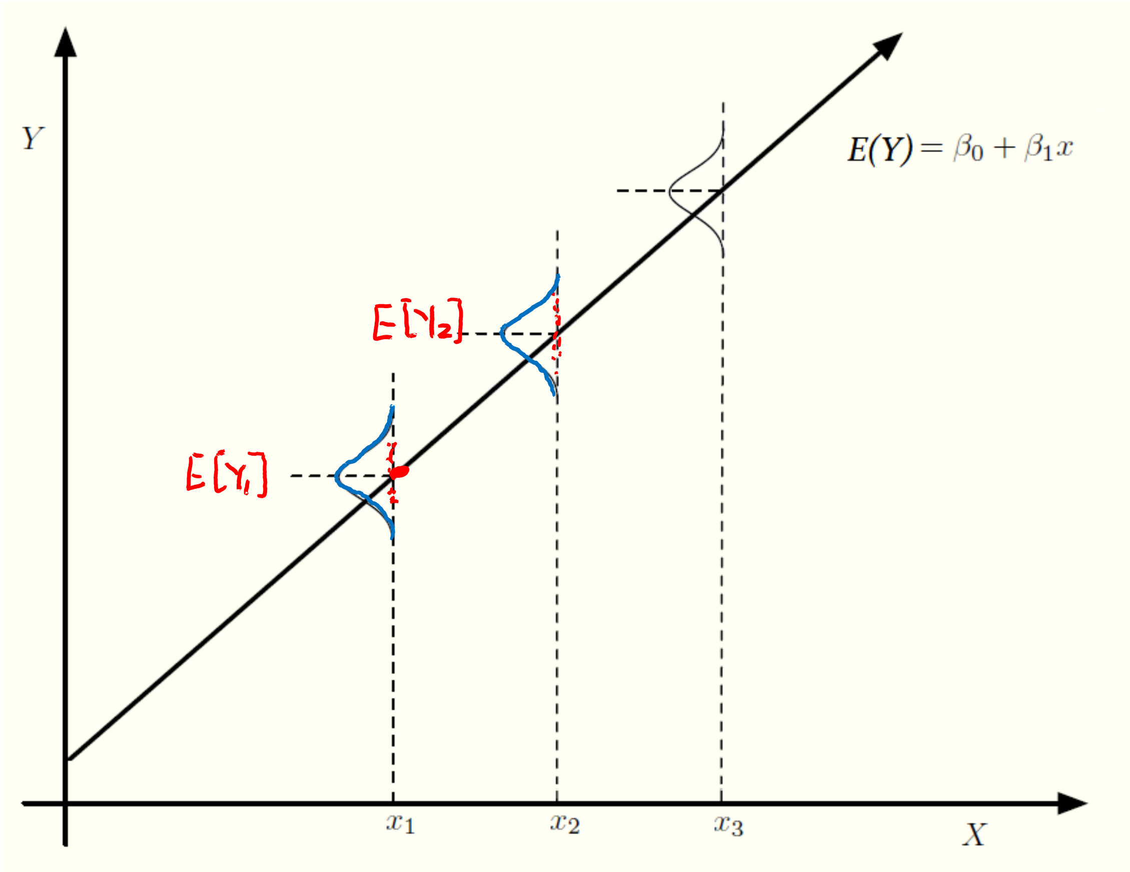
$Y_i = \dfrac{\beta_0 x_i}{\beta_1 + x_i} + \varepsilon_i$

are <u>NOT</u> linear models for Y_i

$E[Y_i] = \beta_0 + \beta_1 x_i$ is a linear function of β_0 and β_1.

e.g. $Y_i = \beta_0 + \beta_1 \underbrace{(\log x_i)}_{d_i} + \varepsilon_i$, $Y_i = \beta_0 + \beta_1 \underbrace{(x_i^2)}_{d_i} + \varepsilon_i$

$SLR: Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

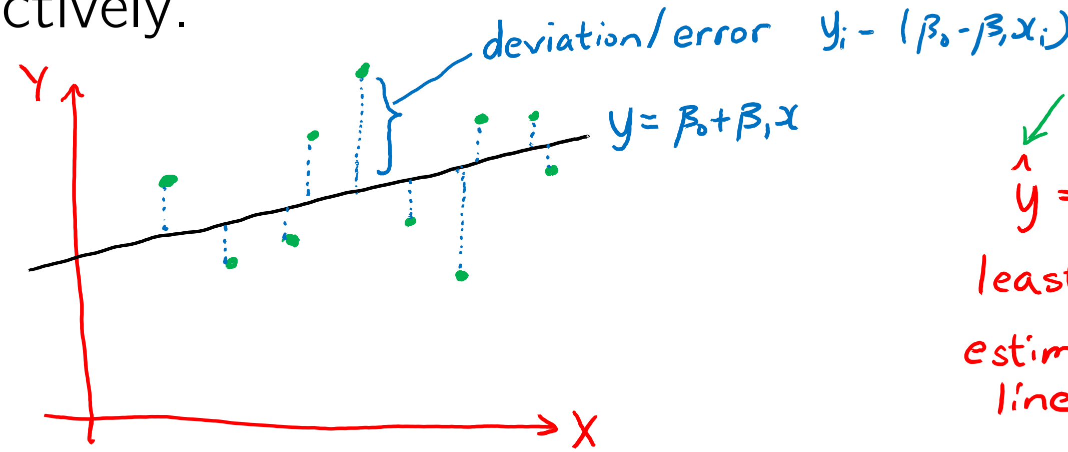


Least square estimation

The least squares estimates of β_0 and β_1 are the values that jointly minimize

$$\boxed{Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 .} = \sum_{i=1}^{n} \varepsilon_i^{2}$$

The least squares estimates of $\underline{\beta_0}$ and $\underline{\beta_1}$ are denoted by $\underline{\hat{\beta}_0}$ and $\underline{\hat{\beta}_1}$ respectively.



predicted value of y

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

least squares line,

estimated regression line

Lemma 5

Suppose $u_1, u_2, \ldots, u_n$ are numbers and let

$$\boxed{q(\gamma) = \sum_{i=1}^{n} (u_i - \gamma)^2 .}$$

Then $q(\gamma)$ is uniquely minimized when $\underline{\gamma = \bar{u}}$.

$q(\gamma) = \sum_{i=1}^{n} (u_i - a_i \gamma)^2$ is uniquely minimised at $\gamma = \dfrac{\sum_{i=1}^{n} a_i u_i}{\sum_{i=1}^{n} a_i^2}$.

If $a_i = 1$, $\gamma = \bar{u}$.

Proof of Lemma 5

$a\gamma^2 + b\gamma + c$

$\gamma = -\frac{b}{2a}$

$$q(\gamma) = \sum_{i=1}^{n} (u_i - a_i\gamma)^2$$

$$= \sum_{i=1}^{n} (u_i^2 - 2u_i a_i \gamma + a_i^2 \gamma^2)$$

$$= \sum_{i=1}^{n} u_i^2 - 2\gamma \sum_{i=1}^{n} a_i u_i + \gamma^2 \sum_{i=1}^{n} a_i^2$$

$$= \underbrace{\left(\sum_{i=1}^{n} a_i^2\right)}_{a} \gamma^2 \underbrace{- 2\left(\sum_{i=1}^{n} a_i u_i\right)}_{b} \gamma + \underbrace{\left(\sum_{i=1}^{n} u_i^2\right)}_{c}$$

This is a quadratic of γ.

The coefficient of γ^2 is non-negative, so the unique minimum occurs at

$$\gamma = -\frac{b}{2a} = \cancel{-}\frac{\cancel{-2}\sum_{i=1}^{n} a_i u_i}{\cancel{2}\sum_{i=1}^{n} a_i^2} = \frac{\sum_{i=1}^{n} a_i u_i}{\sum_{i=1}^{n} a_i^2}.$$

If $a_i = 1$, we have $\gamma = \frac{1}{n}\sum_{i=1}^{n} u_i = \bar{u}$.

Theorem 7

The least square estimates for β_0 and β_1 are given by

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

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

where

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

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

Proof of Theorem 7

$Q(\beta_0, \beta_1)$ can be minimised in two stages.

For β_0: $Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{i=1}^{n} [\underbrace{(y_i - \beta_1 x_i)}_{u_i} - \underbrace{\beta_0}_{\gamma}]^2$

Applying Lemma 5, we have $\hat{\beta}_0 = \bar{u} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_1 x_i)$

$= \bar{y} - \beta_1 \bar{x}$

For β_1: Substitute $\hat{\beta}_0$ for β_0 in $Q(\beta_0, \beta_1)$, then minimise Q with respect to β_1.

$$\begin{aligned} Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \beta_1 x_i)]^2 \\ &= \sum_{i=1}^{n} [y_i - (\bar{y} - \beta_1 \bar{x} + \beta_1 x_i)]^2 \\ &= \sum_{i=1}^{n} [\underbrace{(y_i - \bar{y})}_{u_i} - \underbrace{(x_i - \bar{x})}_{a_i} \beta_1]^2 \end{aligned}$$

Applying Lemma 5, we have $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} a_i u_i}{\sum_{i=1}^{n} a_i^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$.

Remarks for Theorem 7

1. It is possible to also minimize $Q(\beta_0, \beta_1)$ using calculus.

Solve $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$.

$$\left.\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2n(\bar{y} - \beta_0 - \beta_1 \bar{x}) = 0 \\ \frac{\partial Q}{\partial \beta_1} &= -2\left(\sum_{i=1}^{n} x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2\right) = 0 \end{aligned}\right]$$

Least squares equations for SLR

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Remarks for Theorem 7

2. If $S_{xx} = 0$, then the formula for the least squares estimates cannot be evaluated. This has a sensible interpretation because $S_{xx} = 0$ only when all the x-values are identical. In this situation the data clearly contain no information about the slope of the regression line.

3. Note that, by convention, capital letters are used for sum of squares.
   - S_{xx} is called the 'sum of squares due to x'
   - S_{xy} is called the 'cross product sum of squares'

Estimation of σ^2

To estimate σ^2, we use the residual variance

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\underbrace{\left(y_i - \underbrace{\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}_{\hat{y}_i}\right)}_{y_i - \hat{y} \;=\; i^{th} \text{ residual}}{}^2$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the fitted or predicted value for the ith observation.

Estimation of σ^2 (cont.)

The logic of using S_e^2 as an estimator of σ^2 may be understood as follows.

If β_0 and β_1 were known, then

$$\boxed{\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i)} \sim N(0, \sigma^2)$$

and it follows that

$$\sigma^2 = var(\varepsilon_i) = E[(\varepsilon_i - E(\varepsilon_i))^2] = E[(\varepsilon_i - 0)^2] = E[\varepsilon_i^2]$$

$$E\left[\underbrace{\left(Y_i - (\beta_0 + \beta_1 x_i)\right)}_{\varepsilon_i}{}^2\right] = \sigma^2$$

Then

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}\left[\left(Y_i - (\beta_0 + \beta_1 x_i)\right)^2\right]}$$

could be used as an unbiased estimator for σ^2.

Estimation of σ^2 (cont.)

In practice, $\hat{\beta}_0$ and $\hat{\beta}_1$ must be used and the denominator of $n-2$ rather than n is needed to make the estimator unbiased.

Remark:
The three variances discussed to date, namely, S^2 for a single sample, S_p^2 for two independent samples, and S_e^2 for linear regression, have all been constructed by the same principle.

In particular, the appropriate degrees of freedom in each case is given by the rule:

"number of observations" – "number of parameters"