Chapter 13

The Simple Linear Regression Model: Theory

13.1 The model

13.1.1 The data

observations	response variable	explanatory variable
1	y_1	x_1
2	y_2	x_2
:	:	:
n	y_n	x_n

Plotting the data.

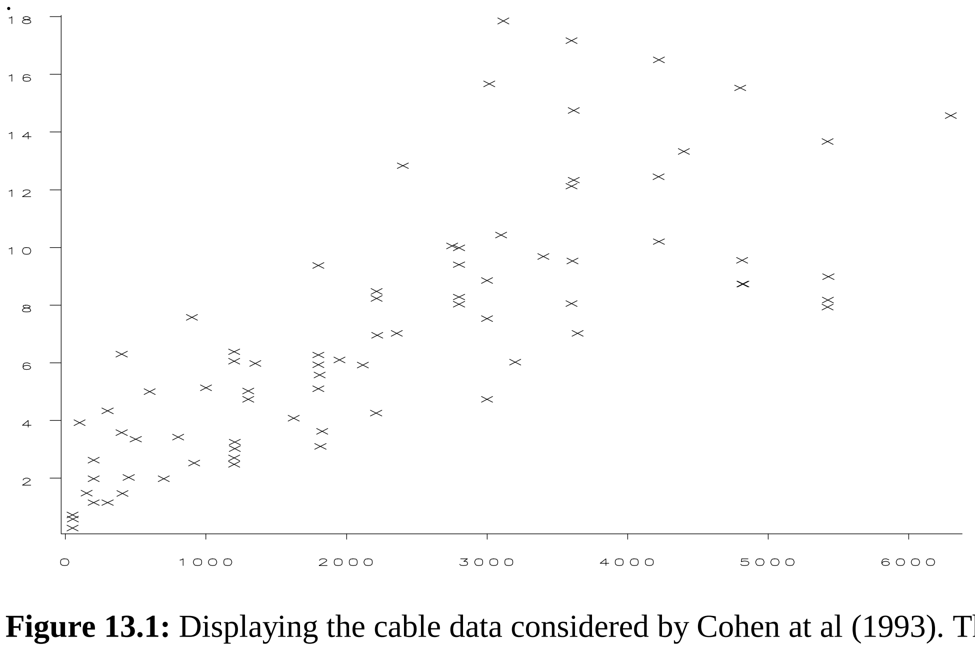

Figure 13.1: Displaying the cable data considered by Cohen at al (1993). There are 79 observations of the number of hours y needed to splice x pairs of wires for a particular type of telephone cable

If the plot is not linear try a simple transformation to linearity. i.e. log, square root, square.

13.1.2 Assumptions for the model

i) The assumption about the linearity of the model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad \text{for} \qquad i = 1, 2, \ldots, n$$

ii) The assumption about the error distribution for ε_i

a) Full distributional assumption for error term ε_i.

$\varepsilon_i \sim N(0, \sigma^2)$ and ε_i and ε_j for $i \neq j$ are independent.

Estimation in this case of the parameters α , β and σ^2 is achieved by Maximum Likelihood.

b) Assumption about the first and second moments of the distribution for ε_i.

$$E(\varepsilon_i) = 0$$

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

$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

Estimation in this case can be achieved by Least Squares.

iii) The assumption about the x-variable.

The *x*-variable is not a random variable and it is fixed at the observed values

13.2 Least squares estimation of parameters

Let $S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$
where the y_i are observed values for the random variable Y_i

In order to find the least square estimators for α and β we need to minimise $S(\alpha, \beta)$ (for fixed y's and x's) with respect to the parameters α and β.

That is we find $\frac{\partial S}{\partial \alpha}$ and $\frac{\partial S}{\partial \beta}$ and we set them equal to zero.

$$\frac{\partial S}{\partial \alpha} = \sum_{i=1}^{n} -2(y_i - \hat{\alpha} - \hat{\beta}\, x_i) = -2\,(\sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta} \sum_{i=1}^{n} x_i) = 0$$

$$\frac{\partial S}{\partial \beta} = \sum_{i=1}^{n} -2x_i\,(y_i - \hat{\alpha} - \hat{\beta}\, x_i) = -2(\sum_{i=1}^{n} x_i y_i - \hat{\alpha} \sum_{i=1}^{n} x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2) = 0$$

with solutions

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

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

The quantities $\hat{y}_i = \hat{\alpha} + \hat{\beta}\, x_i$ are called the **fitted values.**

The quantities $\hat{\varepsilon}_i = y_i - \hat{y}_i$ are called the **residuals.**

13.3 Properties of the least square estimators

Note that both $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the y's . For example for $\hat{\beta}$ we have

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

where $S_{xx} = \sum(x_i - \bar{x})^2$ and $C_i = \dfrac{(x_i - \bar{x})}{S_{xx}}$.

(Prove the above statement for $\hat{\alpha}$).

13.3.1 Expected values for $\hat{\alpha}$ and $\hat{\beta}$

i) $E(\hat{\beta}) = \beta$: $\hat{\beta}$ is an unbiased estimator of β.

Proof

$$E\left(\hat{\beta}\right) = \frac{\sum_{l=1}^{n}(x_i - \bar{x}) \; E(Y_i)}{S_{xx}} \overset{(1)}{=} \frac{\sum_{l=1}^{n}(x_i - \bar{x})(a + \beta x_i)}{S_{xx}}$$

$$\overset{(2)}{=} \frac{\beta \sum(x_i - \bar{x}) \; x_i}{S_{xx}}$$

$$\overset{(3)}{=} \beta$$

(1) since if $y = \sum c_i z_i \Rightarrow E(y) = \sum c_i E(z_i)$

(2) since $\sum(x_i - \bar{x})a = 0$ (prove it)

(3) since $S_{xx} = \sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})x_i$ (prove it)

ii) $E(\hat{\alpha}) = \alpha$: $\hat{\alpha}$ is an unbiased estimator of α.

Proof: $n\hat{\alpha} = \sum_{i=1}^{n} Y_i - \hat{\beta} \sum_{i=1}^{n} x_i$

$$E(n\hat{a})=n\,E(\hat{\alpha})=\sum_{i=1}^{n}E(y_i)-E(\hat{\beta})\sum_{i=1}^{n}x_i$$

$$=\sum_{i=1}^{n}(\alpha+\beta x_i)-\beta\sum x_i$$

$$=n\alpha+\beta\sum x_i-\beta\sum x_i$$

$$=n\,\alpha$$

or

$$E(\hat{\alpha})=\alpha$$

13.3.2 The Variances of $\hat{\alpha}$ and $\hat{\beta}$

i) $$\mathrm{Var}(\hat{\beta})=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sigma^2}{S_{xx}}$$

Proof.

$$\mathrm{Var}(\hat{\beta})=\left\{\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{S_{xx}}\right\}^2\ \mathrm{var}(Y_i)$$

$$=\ \frac{\sigma^2}{S_{xx}}$$

so

$$V\hat{a}r(\hat{\beta})=\frac{\hat{\sigma}^2}{S_{xx}}$$

ii) $$\mathrm{var}(\hat{\alpha})=\sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]$$

Proof.

$$\mathrm{var}(\hat{\alpha})=\mathrm{var}(\bar{y})+\bar{x}^2\,\mathrm{var}(\hat{\beta})-2\bar{x}\,\mathrm{cov}(\bar{y},\hat{\beta})$$

But $\mathrm{cov}(\bar{y},\hat{\beta})=0$ (see Exercise 13.2), so we have

$$\text{var}(\hat{\alpha}) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}}$$
$$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

hence

$$V\hat{a}r(\hat{\alpha}) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

13.3.3 The Gauss-Markoff theorem

The least-squares estimators $\hat{\alpha}$ and $\hat{\beta}$ have minimum variances among all the linear unbiased estimators.

13.3.4 The Normality assumption of $\hat{\alpha}$ and $\hat{\beta}$

Note that if Y is a linear function of normally distributed variables U_i i.e.

$$Y = c_1 U_1 + c_2 U_2$$

Y will be Normally distributed i.e.

$$Y \sim N\left(\mu, \sigma^2\right).$$

The L.S. estimators $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of Y_i which is

$$Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$$

so $\hat{\alpha}$ and $\hat{\beta}$ will be Normally distributed as

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]\right)$$

$$\hat{\beta} \sim N\left(\beta\ ,\ \frac{\sigma^2}{S_{xx}}\right)$$

13.4 Hypothesis testing

13.4.1 Estimation of σ^2

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}$$

where $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ are the fitted values $\hat{\varepsilon}_i = y_i - \hat{y}_i$ the residuals and $n-2$ are the residual degrees of freedom (df).

13.4.2 t-test for $\hat{\beta}$ and $\hat{\alpha}$

Note that if $\left.\begin{array}{l} z \sim N(0,1) \\ \omega \sim \chi^2(\eta) \end{array}\right\}$ then $t = \frac{z}{\sqrt{\frac{\omega}{\eta}}} \sim t(\eta)$

and z and ω are independent we have that

$$\frac{\hat{\beta} - \beta}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

and that

$$\frac{\sum (y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2 (n-2)$$

Also $\hat{\beta}$ and $\sum (y_i - \hat{y})^2$ are independent (not proven). So

$$t = \frac{\dfrac{\dfrac{\hat{\beta} - \beta}{\sigma}}{\sqrt{S_{xx}}}}{\sqrt{\dfrac{\dfrac{\sum (y_i - \hat{y})^2}{\sigma^2}}{n-2}}} = \frac{\hat{\beta} - \beta}{\dfrac{\hat{\sigma}}{\sqrt{S_{xx}}}} = \frac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim t(n-2)$$

where $se(\hat{\beta}) = \dfrac{s}{\sqrt{S_{xx}}} = \dfrac{\hat{\sigma}}{\sqrt{S_{xx}}}$ is the standard error of $\hat{\beta}$.

We can test hypothesis for β using the statistics t.

For example to test

$$Ho : \beta = \beta_o \qquad\qquad H_1 : \beta \neq \beta_o$$

calculate

$$t = \frac{\hat{\beta} - \beta_o}{se(\hat{\beta})}$$

Now if

$$|t| > t_{n-2, \frac{a}{2}}$$

reject the null hypothesis H_0 and accept the alternative H_1, otherwise accept H_0.

Note that a is the *significant level* of the test and not the constant parameter α of the linear model.

To test hypothesis about α i.e.

$$Ho : \alpha = \alpha_o \qquad\qquad H_1 : \alpha \neq \alpha_o$$

use the test statistic

$$t = \frac{\hat{\alpha} - \alpha_o}{se(\hat{\alpha})}.$$

13.4.3 C.I. for α and β

A $(1-a)100\%$ C. I. for β is given by

$$\hat{\beta} \pm t_{n-2,\frac{a}{2}} \times se(\hat{\beta})$$

and for α is given by

$$\hat{\alpha} \pm t_{n-2,\frac{a}{2}} \times se(\hat{\alpha})$$

13.5 Prediction and Confidence Intervals

13.5.1 Confidence Intervals for $\mu_o = a + bx_o$

Note that the expected value for y_0 the value of the y-variable when the explanatory variable is at x_0 is

$$E(y_o) = \mu_o = \alpha + \beta x_o \quad :$$

The fitted value at the point x_0 is defined as

$$\hat{y}_0 = \hat{\mu}_o = \hat{\alpha} + \hat{\beta} x_o = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_0 = \bar{y} + \hat{\beta}(x_0 - \bar{x})$$

with expected values

$$E(\hat{y}_o) = \alpha + \beta x_o = \mu_o \text{ as } E(\hat{\alpha}) = \alpha \text{ and } E(\hat{\beta}) = \beta$$

So the fitted value $\hat{y}_o$ is unbiased for μ_o. The variance for $\hat{y}_o$ is

$$\operatorname{var}(\hat{y}_o) = \operatorname{var}(\bar{y}) + \operatorname{var}(\hat{\beta})(x_0 - \bar{x})^2 + 2(x_0 - \bar{x})\operatorname{cov}(\bar{y}, \hat{\beta})$$
$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}(x_0 - \bar{x})^2 = \sigma^2\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]$$

so an estimate for the variance is given by.

$$\hat{\operatorname{var}}(\hat{y}_o) = s^2\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]$$

where $s^2 = \sum (y_i - \hat{y})^2 / n - 2$.

Since $\hat{y}$ is a linear combination of Normally distributed variables, it is Normally distributed; i.e.

$$\hat{y}_o \sim N\left(\mu_o, \sigma^2\left[\frac{1}{n} + \frac{(x_o - \mu)^2}{\sum (x_i - \bar{x})^2}\right]\right)$$

$$\Rightarrow z_o = \frac{\hat{y}_o - \mu_o}{\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}} \sim N(0,1)$$

or

$$t = \frac{\hat{y}_o - \mu_o}{S^2\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]} \sim t_{n-2}$$

so a C.I for μ_o is given by

$$\hat{y}_o \pm t_{n-2, \frac{a}{2}} \times se(\hat{y}_o)$$

where

$$se(\hat{y}_o) = S\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]^{\frac{1}{2}}$$

13.52 Prediction Interval for y_o^* / x_o, a future observation for y_0.

Let y_o^* / x_o denote a future observation of the y-variable at the x-variable value x_o. Then

$$E(y_0^* / x_o) = \alpha + \beta x_o = \mu_o$$

Since $\hat{y}_o = \hat{\alpha} + \hat{\beta} x_o$ is an unbiased estimator for μ_o it can be used to predict the mean of a future observation y_o^* / x_o.

In general in order to evaluate how good our predictor $\hat{y}$ is for predicting a further observation y^* we have to know the mean square error for prediction or PSE.

Definition: $PSE(y^*) = E(y^* - \hat{y})^2$

Theorem: Let $\hat{y}$ be an estimate of μ and let y^* be a new observation such that $E(y^*) = \mu$. Then $PSE(y^*) = Var(y^*) + MSE(\hat{y})$ where $MSE(\hat{y}) = E(\hat{y} - \mu)^2$.

Proof:

$$\begin{aligned} PSE(y^*) &= E(y^* - \hat{y})^2 \\ &= E\left[(y^* - \mu) - (\hat{y} - \mu)\right]^2 \\ &= E\left[(y^* - \mu)^2 - 2(y^* - \mu)(\hat{y} - \mu) + (\hat{y} - \mu)^2\right] \end{aligned}$$

y^* is independent of $\hat{y}$, as y^* is a new observation. Hence

$$E\left[(y^* - \mu)(\hat{y} - \mu)\right] = E(y^* - \mu)E(\hat{y} - \mu) = 0 \text{ as } E(y^*) = \mu$$

Hence

$$\begin{aligned} PSE(y^*) &= E\left[(y^* - \mu)^2 + E(\hat{y} - \mu)^2\right] \\ &= Var(y^*) + MSE(\hat{y}) \\ &= Var(y^*) + Var(\hat{y}) + (bias)^2 \end{aligned}$$

In the simple linear regression example we have

$$E\left(y_o^* - \hat{y}_o\right)^2 = Var\left(y_o^*\right) + Var\left(\hat{y}_o\right) \qquad \text{since } \hat{y}_o \text{ is unbiased}$$

$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{\left(x_o - \bar{x}\right)^2}{\sum \left(x_i - \bar{x}\right)^2}\right]$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{\left(x_o - \bar{x}\right)^2}{\sum \left(x_i - \bar{x}\right)^2}\right]$$

A $100(1-a)\%$ prediction interval for y_0^* / x_o is given by

$$\hat{y}_o \pm t_{n-2,\frac{a}{2}}\; s\left[1 + \frac{1}{n} + \frac{\left(x_o - \bar{x}\right)^2}{\sum \left(x_1 - \bar{x}\right)^2}\right]^{\frac{1}{2}}$$

13.6 Maximum likelihood estimation of the parameters α, β and σ^2 in the simple linear regression.

The **likelihood function** is the probability of observing the sample seeing as a function of the parameter rather than a function of the random variables.

For independent random variables $x_1, x_2 \ldots x_n$ the likelihood will be

$$L(\theta) = f\left(x_1;\theta\right) f\left(x_2;\theta\right) \ldots f\left(x_n;\theta\right)$$

In the simple regression model we have

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where

$$\varepsilon_i \overset{ind}{\sim} N(0,\sigma^2) \quad \Rightarrow y_i \overset{ind}{\sim} N\left(\alpha + b x_i, \sigma^2\right)$$

i.e.

$$E\left(Y_i\right) = \alpha + \beta x_i$$

$$\text{var}(Y_i) = \sigma^2$$

and $\theta = (\alpha, \beta, \sigma^2)$.

The likelihood for one observation is

$$L(\alpha, b, \sigma^2; y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \alpha - bx_i)^2 \right\}$$

for n independent observations the likelihood will be

$$L(\alpha, b, \sigma^2 / y_1 \ldots y_n) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \alpha - bx_i)^2 \right\}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - bx_i)^2 \right\}.$$

Note that $S = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ is the function that we minimised in the least square estimation approach.

In order to find the MLE's for α, β and σ^2 we have to maximise $L(\alpha, \beta, \sigma^2)$ with respect to the parameters or equivalently maximise $\log L(\alpha, b, \sigma^2) = \ell(\alpha, \beta, \sigma^2)$
Now

$$\ell(\alpha, \beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\pi\sigma^2} \sum_{1}^{n} (y_i - \alpha - \beta x_i)^2$$

so we differentiate with respect to α, β and σ^2

$$\frac{\partial \ell}{\partial \alpha} = \frac{1}{\sigma^2} \sum_{1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right) = 0$$

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2} \sum_{1}^{n} \left(x_i \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right) \right) = 0$$

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 = 0$$

solving for $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ we have

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{\sum \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2}{n}$$

Note:

i) $\hat{\alpha}$ and $\hat{\beta}$ are also the least-square estimators. This is because the maximisation of the log-likelihood (for fixed σ^2) is the equivalent of the minimisation of the least-square quantity $S = \sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)^2$.

ii) We generally prefer to use an unbiased estimator of σ^2 given by

$$s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2}{n-2} = \frac{D}{df} \quad \leftarrow \text{ deviance}$$

$\uparrow$ the residual degrees of freedom

Exercise 13.1: Simple linear regression theory

a) Consider the simple linear regression model of the form

$$Y_i = a + bx_i + \varepsilon_i \qquad \text{for } i = 1,2\ldots,n$$

where
Y_i is the response variable,

x_i is the independent variable,
a and b are parameters to be estimated.
ε_i , for $i = 1,2...,n$ are independent Normally distributed variables with mean 0 and variance σ^2.

i) Find the likelihood function for a single observation and hence show that the log-likelihood function for all n observations from the above model is

$$l(a,b,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

ii) Give the Normal equations used to find the Maximum likelihood estimators for the parameters a , b and σ^2, and state the resulting maximum likelihood estimators of a , b and σ^2.

b)

i) State the distribution of $\hat{b}$ and $\frac{\sum_{i=1}^{n}\hat{e}_i^2}{\sigma^2}$ where $\hat{e}_i = y_i - \hat{a} - \hat{b}x_i$, i.e. the residual for the *i*th observation and the numerator in the expression is the Residual Sum of Squares (RSS). Note that the variance of $\hat{b}$ is $\frac{\sigma^2}{S_{xx}}$ where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

ii) We know that $\hat{b}$ and $\sum_{i-1}^{n}\hat{e}_i^2$ are independent.
We also know that if $z \sim N(0,1)$ and $w \sim \chi_\nu^2$, independently,
then $\frac{z}{\sqrt{(w/\nu)}} \sim t_\nu$.
Use this to construct a 100(1-α)% confidence interval for b.

Exercise 13.2: Simple linear regression theory

For a simple linear regression model, prove that $cov(\bar{y}, \hat{\beta}) = 0$

Note: an outlined method is as follows:

$$
\begin{aligned}
cov(\bar{y}, \hat{\beta}) &= E\left[(\bar{y}-\mu)(\hat{\beta}-\beta)\right] \\
&= E\left[(\bar{y}-\mu)\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\beta\right)\right] \\
&= E\left[(\bar{y}-\mu)\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\beta\right)\right] - E\left[(\bar{y}-\mu)\beta\right] \\
&= E\left[(\bar{y}-\mu)\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\beta\right)\right] - 0 \qquad \text{(Why?)}
\end{aligned}
$$

Now

$$
\begin{aligned}
E\left[(y_i-\bar{y})(\bar{y}-\mu)\right] &= E\left[\left((n-1)y_i - y_1 - \ldots - y_{i-1} - y_{i+1} - \ldots y_n\right)(y_1 + \ldots + y_n - n\mu)/n^2\right] \\
&= E\left[\left((n-1)\{y_i-\mu\} - \{y_i-\mu\} - \ldots\right)\left(\{y_1-\mu\} + \ldots + \{y_1-\mu\}\right)/n^2\right] \\
&= (n-1)\sigma^2 - \sigma^2 - \sigma^2 \ldots - \sigma^2 \\
&= 0
\end{aligned}
$$

as

$$E\{(y_i-\mu)(y_j-\mu)\} = 0 \qquad \text{(Why?)}$$

Hence deduce $cov(\bar{y}, \hat{\beta}) = 0$.

Exercise 13.3: Simple linear regression theory

For the regression $y_i = a + bx_i + \varepsilon_i$, $\varepsilon_i \sim N(0,1)$ show that the least squares estimate of b is given by

$$\hat{b} = \frac{S_{xy} - n\overline{x}\overline{y}}{S_{xx} - n\overline{x}^2} .$$

A test of b= 0 can be based upon

$$T = S_{xy} - n\overline{y}\overline{x} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \overline{x}$$
$$= \sum_{i=1}^{n} y_i \left(x_i - \overline{x}\right)$$

Show that, under $H_0 : b = 0$, we have $E(T) = 0$.

Also show that $V(T) = \sigma^2 \sum \left(x_i - \overline{x}\right)^2$.

Deduce that $\dfrac{S_{xy} - n\overline{y}\overline{x}}{\sigma\sqrt{\sum \left(x_i - \overline{x}\right)^2}} \sim N(0,1)$

And hence that $\dfrac{S_{xy} - n\overline{y}\overline{x}}{s\sqrt{\sum \left(x_i - \overline{x}\right)^2}} \sim t_{n-1}$

Practical 13: Simple Linear Regression

The data set in the file SHARED (K):\SCTMS\SOM\MA2010\REGRESSION\FOOT GESTATION TIME.SAV comprises measurements of foetal foot length in mm (Y) and gestational age in weeks (X) for 450 foetuses.

1. Produce a scatter plot of Y against X using the procedure
   > Graphs > Scatter > Simple Scatter > Define | Y Axis ‘foot’ | X Axis ‘gest’
   > OK.
   Comment on this plot.

2. Fit a simple linear regression line using
   > Analyse > Regression > Linear | Dependent: ‘foot’ | Independent(s): ‘gest’ to declare your y and x variable and fit the model.
   You can use the PLOTS option to get the residual plots.
   > Plots | Y: ZRESID | X: ZPRED | ✓ Histogram | ✓ Normal probability plot
   > Continue > OK

   i) State the model fitted and its parameter estimates. Interpret these estimates.
   ii) Test whether there is a linear relationship between foot length and gestational age.
   iii) State the assumptions necessary for your model to be valid.
   iv) Do the residual plots show that any of the assumptions does not hold?