MAS programmes - Statistical Analysis (Autumn Term)

1 Review of Basic Statistical Concepts

1.1 Random samples and their descriptive statistics

Let $y_1, y_2, \ldots, y_n$ denote a sample of size n. Typically, these will be used to refer to unrealized random variables: however, where the context makes it obvious, they will instead be used to refer to actual, realized, observations. (If there is any danger that the usage may be unclear, then upper case symbols will be used to refer to the random variables and lower case ones will be used to refer to their observed realizations).

We may calculate statistics to summarize the data. For example, the *sample mean* $\bar{y}$ is given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and the *sample variance* s^2 by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The *sample standard deviation* s is the square root of the sample variance.

The *order statistics* $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ are the sample data written in increasing order, so that

$$y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}.$$

In particular, $y_{(1)}$ is the *minimum* and $y_{(n)}$ the *maximum* of the sample data. The *sample median* is a statistic such that half of the sample data lie below it and half above. More specifically, if the sample size n is an even number, $n = 2m$, say, then the sample median is given by

$$\frac{y_{(m)} + y_{(m+1)}}{2}.$$

If the sample size is an odd number, $n = 2m - 1$, say, then the median is given by $y_{(m)}$.

The sample mean and median are both measures of location, giving what may be regarded as a typical sample value. The median has the advantage that it is less sensitive to extreme values in the sample data.

The *lower quartile* is a number such that one quarter of the sample data are 'smaller' than it and three quarters larger. The *upper quartile* is a number such that three quarters of the sample data are smaller than it and one quarter 'larger'.

A *boxplot* can be a useful means of displaying data, the points of the boxplot relating to the measures given from the order statistics above.

We now return to the general case by assuming that the data are a *random sample* from some infinite population with population mean μ and population variance σ^2.

Definition 1.1 (Random Sample)
The sample data $y_1, y_2, \ldots, y_n$ are a random sample if they are independently and identically distributed random variables (i.i.d. r.v.'s).

In such a situation it is always the case that the sample mean $\bar{y}$ is an unbiased estimator of the population mean μ and the sample variance s^2 is an unbiased estimator of the population variance σ^2 (see Prop. 1.2). Expressing this mathematically,

$$\mathsf{E}[\bar{y}] = \mu,$$

$$\mathsf{E}[s^2] = \sigma^2.$$

Furthermore,

$$\mathsf{Var}(\bar{y}) = \frac{\sigma^2}{n}.$$

Using an alternative phraseology, we say that the *sampling distribution* of the sample mean $\bar{y}$ has mean μ and variance σ^2/n. Now σ^2 is generally unknown, but, replacing it by the unbiased estimator s^2, we see that an unbiased estimator of $\mathsf{Var}(\bar{y})$ is given by s^2/n. The square root of this quantity is often referred to as the '*standard error of the mean*',

$$\frac{s}{\sqrt{n}}.$$

Proposition 1.2 ($\overline{y}$ and s^2 are unbiased estimators for μ and σ^2, respectively)
Suppose $y_1, y_2, \ldots, y_n$ is a random sample where $\mathsf{E}[y_i] = \mu$ and $\mathsf{Var}(y_i) = \sigma^2$, $i = 1, \ldots, n$. Then
(i) $\mathsf{E}[\overline{y}] = \mu$
(ii) $\mathsf{E}[s^2] = \sigma^2$.

Proof
(i)

$$\mathsf{E}[\overline{y}] = \mathsf{E}[\frac{1}{n}\sum_{i=1}^{n} y_i] = \frac{1}{n}\mathsf{E}[\sum_{i=1}^{n} y_i] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[y_i] = \frac{1}{n}n \times \mu = \mu.$$

(ii) Observe that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i^2 - 2y_i\overline{y} + \overline{y}^2)$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} y_i^2 - 2\overline{y}\sum_{i=1}^{n} y_i + n\overline{y}^2\right\} = \frac{1}{n-1}\left\{\sum_{i=1}^{n} y_i^2 - 2n\overline{y}^2 + n\overline{y}^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right\} = \frac{1}{n-1}\left\{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2\right\}.$$

So

$$\mathsf{E}[s^2] = E\left[\frac{1}{n-1}\left\{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right\}\right] = \frac{1}{n-1}\sum_{i=1}^{n}\mathsf{E}[y_i^2] - \frac{n}{n-1}\mathsf{E}[\overline{y}^2].$$

But

$$\mathsf{E}[y_i^2] = \mathsf{Var}(y_i) + \mathsf{E}[y_i]^2 = \sigma^2 + \mu^2$$
$$\mathsf{E}[\overline{y}^2] = \mathsf{Var}(\overline{y}) + \mathsf{E}[\overline{y}]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Hence

$$\mathsf{E}[s^2] = \frac{n}{n-1}(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2.$$

Remarks 1.3
(i) Note that the estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$ is a biased estimator for σ^2. In fact, it slightly underestimates σ^2, because

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

therefore

$$\mathsf{E}[\hat{\sigma}^2] = \frac{n-1}{n}\mathsf{E}[s^2] = \frac{n-1}{n}\sigma^2.$$

However, $\hat{\sigma}^2$ is asymptotically unbiased as $n \longrightarrow \infty$, i.e. $\mathsf{E}[\hat{\sigma}^2] \longrightarrow \sigma^2$.

(ii) $\overline{y}$ and y_1 are both unbiased estimators of μ; however, we might expect $\overline{y}$ to give better performance since it takes into account more of the available information. In fact, we say that $\overline{y}$ is more efficient than y_1 as an estimator of μ because

$$\mathsf{Var}(\overline{y}) = \frac{\sigma^2}{n} < \sigma^2 = \mathsf{Var}(y_1)$$

for $n > 1$, i.e. $\overline{y}$ is a 'less variable' statistic.

(iii) If normality of the observations can be assumed, then $\hat{\sigma}^2$ noted above is the *maximum likelihood estimator* (MLE) of σ, and $\bar{y}$ is the MLE of μ. It can further be shown, under the assumption of normality, that $\bar{y}$ and s^2 are *independent*.

1.2 The normality assumption

In the development of many basic statistical techniques, it is assumed that sample data are a random sample from an $N(\mu, \sigma^2)$ distribution: the population distribution of the random variable under consideration is normal with some mean μ and variance σ^2, where μ and σ^2 are generally unknown. It then follows that the sampling distribution of the sample mean $\bar{y}$ is also normal:

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Hence, using the properties of the normal distributions,

$$\frac{\sqrt{n}(\bar{y} - \mu)}{\sigma} \sim N(0, 1),$$

the standard normal distribution.

Normality of data can be assessed by looking at a boxplot or histogram (to assess symmetry), or using a *normal probability plot* (or quantile-quantile plot). In such a plot the ordered observations $y_{(i)}$ are plotted against quantiles of the standard normal distribution,

$$\Phi^{-1}\left(\frac{i-\frac{1}{2}}{n}\right) \quad i = 1, \ldots, n \quad [\text{if } n \geq 11],$$

where Φ is the cumulative distribution function of the standard normal distribution. If the observations are approximately normally distributed then the plot should be approximately linear.

Because in general σ is unknown, it turns out that, in making inferences about μ, the population standard deviation σ in the above expression has to be replaced by the sample standard deviation s. The corresponding sampling distribution is:

$$\frac{\sqrt{n}(\bar{y}-\mu)}{s} \sim t_{n-1}, \tag{1.1}$$

a (Student's) *t-distribution* with $n-1$ degrees of freedom.

We shall need to make use of what are known as the percentage points of the t-distribution. The $100\alpha\%$ *(percentage) point* of the t_{n-1} distribution is the number $t_{n-1;\alpha}$ such that if $T \sim t_{n-1}$ then

$$\mathbb{P}(T > t_{n-1;\alpha}) = \alpha.$$

These percentage points are numbers that may be looked up in t-tables, or calculated in a software package (such as R).

The problems of inference that we shall be dealing with may be divided into those of (a) estimation and (b) hypothesis testing. The obvious *point estimate* of μ is the sample mean $\bar{y}$, but we should also provide an *interval estimate*, that is, a confidence interval.

From Equation (1.1) and the definition of the percentage points,

$$\mathbb{P}\left(-t_{n-1;\alpha/2} < \frac{\sqrt{n}(\bar{y}-\mu)}{s} < t_{n-1;\alpha/2}\right) = 1-\alpha. \tag{1.2}$$

Rearranging the left hand side of Equation (1.2), we obtain

$$\mathbb{P}\left(\bar{y} - t_{n-1;\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha.$$

We say that

$$\left(\bar{y} - t_{n-1;\alpha/2}\frac{s}{\sqrt{n}} \;,\; \bar{y} + t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ (equal tails) *confidence interval* for μ.

- Recall that $s/\sqrt{n}$ is the standard error of the mean.

We may wish to test the *null hypothesis* that the population mean is equal to some particular value μ_0 against the *alternative hypothesis* that it is not equal to μ_0. Thus we test

$$H_0 : \mu = \mu_0$$

against

$$H_1 : \mu \neq \mu_0.$$

The test procedure is based upon the test statistic derived from Equation (1.1),

$$t = \frac{\sqrt{n}(\bar{y} - \mu_0)}{s},$$

which, under H_0, is known to have the t_{n-1} distribution. Broadly speaking, we reject H_0 if the absolute value of the test statistic is large enough. More specifically, we can calculate the *p-value* (or *significance level*) p of the test statistic, which is the probability under H_0 of obtaining the observed value of the test statistic or a more extreme one.

The p-value expresses the weight of evidence against H_0. The smaller the p-value, the stronger is the evidence against H_0: the usual benchmark values of p are 0.05, 0.01 ... Using the traditional phraseology, we say that "we reject H_0 at the $100\alpha\%$ significance level" if $p < \alpha$ where, conventionally, $\alpha = 0.05,\ 0.01 \ldots$

- We shall usually assume a two-sided alternative hypothesis $H_1 : \mu \neq \mu_0$, but we could use a one-sided alternative, $H_1 : \mu < \mu_0$ or $H_1 : \mu > \mu_0$, which would affect the calculation of the p-value.

In making inferences about σ^2, the relevant result in distribution theory is:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}, \tag{1.3}$$

a *chi-square distribution* with $n-1$ degrees of freedom. Thus, for example,

$$\mathbb{P}\left(\chi^2_{n-1;1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1;\alpha/2}\right) = 1 - \alpha.$$

Hence

$$\mathbb{P}\left(\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}\right) = 1 - \alpha.$$

and

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

The following facts about chi-square distributions will be useful.

1. If $U \sim \chi^2_\nu$ then $\mathrm{E}(U) = \nu$ and $\mathrm{Var}(U) = 2\nu$.

2. If $U_1, U_2, \ldots, U_k$ are independently distributed random variables, $U_i \sim \chi^2_{\nu_i}$ $i = 1, \ldots, k$, then $U_1 + U_2 + \ldots + U_k \sim \chi^2_\nu$, where $\nu = \sum_{i=1}^k \nu_i$.

3. If U and V are a pair of independent random variables, with $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ then the random variable F defined by

$$F = \frac{U/\nu_1}{V/\nu_2}$$

has what is known as the *(Fisher's) F-distribution with ν_1 and ν_2 degrees of freedom* (the F_{ν_1,ν_2} distribution). Note that $\mathrm{E}(U/\nu_1) = 1$ and $\mathrm{E}(V/\nu_2) = 1$. Although $\mathrm{E}(F) \neq 1$, the distribution of F is, roughly speaking, centred around 1. In fact, for $v_2 > 2$ we have,

$$\mathrm{E}(F) = \frac{v_2}{v_2 - 2}$$

Also recall the relationship between the F and t distributions, *viz*

$$T \sim t_m \Longleftrightarrow T^2 \sim F_{1,m}$$

Example: Consider the running times measured in minutes of a group of 40 runners for a certain cross country run.

Running Times (mins)

10.10	12.11	15.15	16.05	11.67	11.54	11.86	15.90	12.35	11.65
12.33	10.23	11.88	11.75	11.94	12.04	13.66	13.45	10.57	10.75
12.43	12.60	12.75	13.00	13.15	13.75	14.00	14.56	12.91	13.45
14.72	12.50	13.25	14.60	14.88	13.23	12.09	13.07	14.08	14.92

The data can be represented graphically as follows:
so that it would appear reasonable to assume normality. Additionally, the sample mean and sample variance are 12.923 minutes and 2.1337 minutes^2 respectively. The standard error of the mean is

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{2.1337}{40}} = 0.2310.$$

A 95% confidence interval for the mean running time is then $12.923 \pm 0.231\,t$, where t is the $2\frac{1}{2}$% point (or 0.975-quantile) from the t_{39} distribution, i.e. $t_{39;0.025} = 2.023$. Hence the confidence interval is (12.46,13.39).

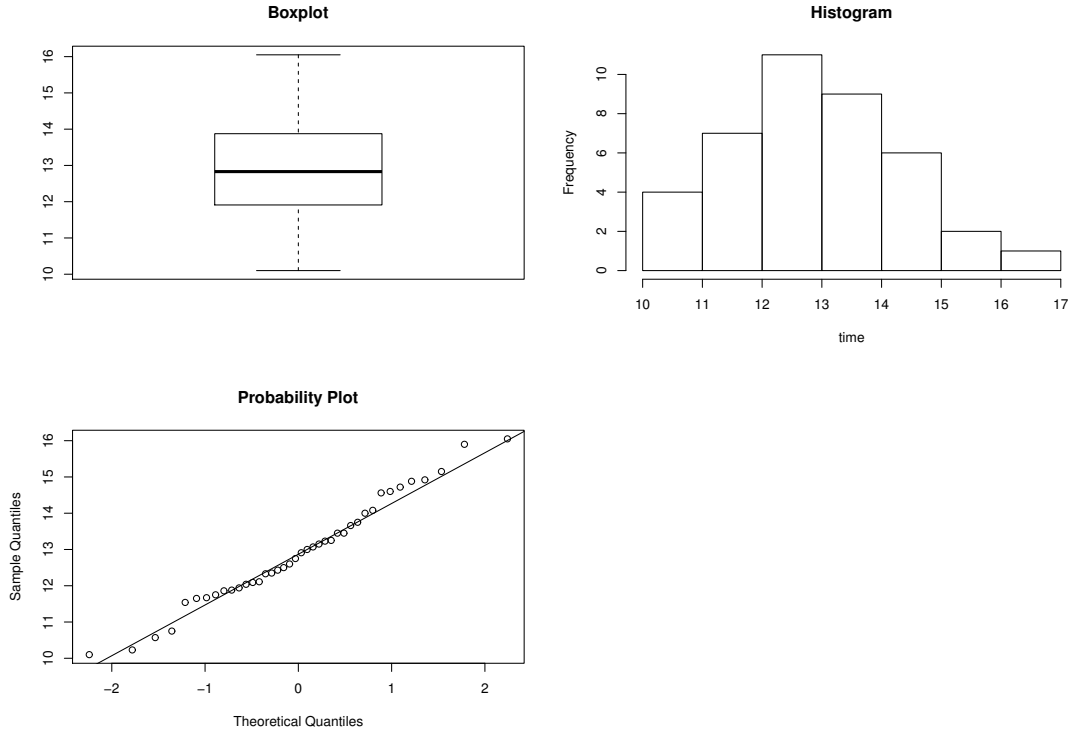
Now, suppose we had wished to test the hypothesis that the running time is not 12.5. i.e. We test $H_0 : \mu = 12.5$.

Informally, we cannot reject such a hypothesis (at the 5% level) since 12.5 is contained in the 95% CI. More formally however, the test statistic is

$$t = \frac{12.923 - 12.5}{0.231} = \frac{0.423}{0.231} = 1.831$$

which must be compared to the t_{39} distribution. The significance probability is

$$2 \times \mathbb{P}(T > 1.831) = 2 \times 0.0374 = 0.0747.$$



1.3 Simple Linear Regression: a review

(Linear regression with one explanatory variable)

Let Y represent a *response* variable of interest and X a single *explanatory* variable. A linear model connecting Y and X takes the form

$$Y = \alpha + \beta X$$

where α and β represent the intercept and slope respectively, i.e. β measures the corresponding change in Y for a unit change in X.

Now suppose that we have n observed pairs (x_i, y_i), $i = 1, \ldots, n$, from which to estimate the *parameters* α and β. The simple linear regression model takes the form

$$y_i = \underbrace{\alpha + \beta x_i}_{\text{systematic part}} + \underbrace{\epsilon_i}_{\substack{\text{random error} \\ \text{term}}} \qquad i = 1, \ldots, n \tag{1.4}$$

where the ϵ_i are independently and identically distributed with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$, i.e. $\epsilon_i \sim$ i.i.d$(0, \sigma^2)$. It follows that the simple linear regression model may be written

$$\mathrm{E}(Y_i) = \mathrm{E}(Y_i | X_i = x_i) = \mu_i = \alpha + \beta x_i \quad \text{and} \quad \mathrm{Var}(Y_i) = \sigma^2$$

so that the response value of each random variable Y_i is *conditional* on the corresponding (given) value of X_i.

1.3.1 Estimation

Least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are chosen to minimize the functional

$$\mathcal{L}(\alpha, \beta) = \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2$$

where the minimized value of $\mathcal{L}(\alpha, \beta)$ is the residual sum of squares, SS_R, i.e. the sum of squared vertical deviations from the fitted line.

It can be shown that

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{1.5}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \tag{1.6}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the observed values of x and y. It is easily seen that the fitted line passes through the point $(\bar{x}, \bar{y})$.

Note that in the above no assumptions are made about the distribution of the errors, and hence the Y_i. The usual assumption is that of normality, so that $\epsilon_i \sim \text{N.I.D}(0, \sigma^2)$ and hence the simple linear regression model can be summarized as

$$Y_i \sim \text{N}(\alpha + \beta x_i, \sigma^2)$$

Another approach to the estimation of parameters is via *maximum likelihood*. Recall that for $Y \sim \text{N}(\mu, \sigma^2)$ with probability density function $f(y) = \{2\pi\sigma^2\}^{-\frac{1}{2}}\text{exp}\{-\frac{1}{2}[(y-\mu)/\sigma]^2\}$, then given a sample (x_i, y_i), $i = 1, \ldots, n$, we have the likelihood expression

$$L(\alpha, \beta; y_1, \ldots, y_n) = \{2\pi\sigma^2\}^{-\frac{n}{2}} \prod_{i=1}^{n} \text{exp}\{-\frac{1}{2}[(y_i - \{\alpha + \beta x_i\})/\sigma]^2\}$$

with corresponding log-likelihood

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

The log-likelihood is therefore maximized, when we minimize the quantity

$$\sum_{i=1}^{n}\{y_i - (\alpha + \beta x_i)\}^2$$

as before. That is, using the method of maximum likelihood leads to exactly the same estimates of α and β as ordinary least squares. Of course, we would also like to have an estimate of σ^2, the *error* or *residual* variance and this is given as

$$s^2 = \frac{SS_R}{n-2} = \frac{\sum_{i=1}^{n}\{y_i - \hat{y}_i\}^2}{n-2} = \frac{\sum_{i=1}^{n}\{y_i - (\hat{\alpha} + \hat{\beta}x_i)\}^2}{n-2}.$$

Note that under the assumption of normality, we have

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}.$$

1.3.2 Inference

If a normal distribution is assumed for Y_i, then the ordinary least squares estimators of the regression coefficients, α and β are themselves normally distributed. In particular, we have

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{s_{xx}}\right) \tag{1.7}$$

so that a test for the value of the slope is easily constructed as a t-test from the following quantity:

$$\frac{\hat{\beta} - \beta}{s/\sqrt{s_{xx}}} \sim t_{n-2}. \tag{1.8}$$

1.3.3 Prediction

A natural use of a regression line is to make predictions of the response at a given level of the explanatory variable, $X = x_0$, say. This is easily predicted using the fitted equation, e.g.

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0$$

A $100(1-\delta)\%$ confidence interval for the expected value of the response variable when the value of the explanatory variable is $x = x_0$ is given by

$$\hat{\alpha} + \hat{\beta}x_0 \pm [t_{n-2,\delta/2}]s\sqrt{\frac{(x_0 - \bar{x})^2}{s_{xx}} + \frac{1}{n}}$$

and the corresponding $100(1-\delta)\%$ prediction interval for the response variable is

$$\hat{\alpha} + \hat{\beta}x_0 \pm [t_{n-2,\delta/2}]s\sqrt{\frac{(x_0 - \bar{x})^2}{s_{xx}} + \frac{1}{n} + 1}$$

Note that for values of x_0 close to the mean $\bar{x}$, the value of $(x_0 - \bar{x})^2$ is small, so that both intervals will be narrower than for x_0 a long way from $\bar{x}$.

The confidence interval takes account of the fact that the regression line is unknown, but merely estimated, so that there is uncertainty about its true position, and hence the average position of Y_0. The prediction interval is wider as it takes account of the further uncertainty about the actual position of Y_0 relative to the line.

1.3.4 Correlation

The (Pearson product-moment) correlation coefficient is one way of quantifying the strength of the linear relationship between variables. It is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where $-1 \leq r \leq 1$. Recall that correlation does not imply causation.

1.3.5 Diagnostics

The fit of the simple linear regression model can be assessed by graphical examination of the residuals calculated using the fitted regression line. The model assumptions are that the error terms are independently and normally distributed with zero mean and constant variance and so the residuals should be consistent with that.