MTH5120 Statistical Modelling 1

Lecture Notes, Semester B 2024

Chris Sutton FIA FHEA, Senior Lecturer in Actuarial Science
School of Mathematical Sciences, Queen Mary University of London

1. Principles of Statistical Modelling

1.1 Why and how models are used

A model is an imitation of a real-world system or process. Models of many activities can be developed, for example, in economics, medicine and business. Suppose we wished to 'predict' the effect that a real-world change would have. In some cases, it might be too risky, or too expensive or too slow, to try a proposed change in the real-world even on a sample basis. Trying out the change first without the benefit of a model could have serious consequences. A model enables the possible consequences to be investigated. The effect of changing certain input parameters can be studied before a decision is made to implement the plans in the real-world.

To build a model of a system or process, a set of mathematical or logical assumptions about how it works needs to be developed. The complexity of a model is determined by the complexity of the relationships between the various model parameters. For example, in modelling the profitability of a business, consideration must be given to issues such as regulations, taxation and sales terms. Future events affecting interest rates, inflation, new business and expenses also affect these relationships.

In order to produce the model and determine suitable parameters, data is needed, and judgements need to be made as to the relevance of the observed data to the future environment. Such data may result from past observations, from current observations or from expectations of future changes.

Where observed data is considered to be suitable for producing the parameters for a chosen model, statistical methods can be used to fit the data.

Before finalising the choice of model and parameters, it is important to consider the objectives for creation and use of the model. For example, in many cases there may not be a desire to create the most accurate model, but instead to create a model that will not understate costs or other risks that may be involved.

While in reality a modelling process does not follow a rigid pattern of prescribed steps, it is helpful in introducing the topic to imagine a set of key steps. In practice, statisticians who build and use models move back and forth between these key steps continuously to improve the model.

The key steps in a modelling process can be described as follows:

i. Develop a well-defined set of objectives which need to be met by the modelling process.
ii. Plan the modelling process and how the model will be validated.
iii. Collect and analyse the necessary data for the model.

iv. Define the parameters for the model and consider appropriate parameter values.
v. Define the model initially by capturing the essence of the real-world system. Refining the level of detail in the model can come at a later stage.
vi. Involve experts on the real-world system you are trying to imitate to get feedback on the validity of the conceptual model.
vii. Write the computer program for the model.
viii. Test the reasonableness of the output from the model.
ix. Review and carefully consider the appropriateness of the model in the light of small changes in input parameters.
x. Analyse the output from the model.
xi. Ensure that any relevant professional guidance has been complied with.
xii. Communicate and document the results and the model.

1.2 Modelling the benefits and limitations

In many areas of work, one of the most important benefits of modelling is that systems with long time frames can be studied in compressed time.

Other benefits include:

• Complex systems with stochastic elements, such as the operation of a company can be studied.

• Different future policies or possible actions can be compared to see which best suits the requirements or constraints of a user.

• In a model of a complex system we can usually get control over the experimental conditions so that we can reduce the variance of the results output from the model without upsetting their mean values.

However, models are not the simple solution to all problems – they have drawbacks that must be understood when interpreting the output from a model and communicating the results.

The drawbacks include:

• Model development requires a considerable investment of time, and expertise. The financial costs of development can be quite large given the need to check the validity of the model's assumptions, the computer code, the reasonableness of results and the way in which results can be interpreted in plain language by the target audience.

• In a stochastic model, for any given set of inputs each run gives only estimates of a model's outputs. So, to study the outputs for any given set of inputs, several independent runs of the model are needed. As a rule, models are more useful for comparing the results of input variations than for optimising outputs.

• Models can look impressive when run on a computer so that there is a danger that one gets lulled into a false sense of confidence. If a model has not passed the tests of validity and verification, its impressive output is a poor substitute for its ability to imitate its corresponding real-world system.

• Models rely heavily on the data input. If the data quality is poor or lacks credibility, then the output from the model is likely to be flawed.

• It is important that the users of the model understand the model and the uses to which it can be safely put. There is a danger of using a model from which it is assumed that all results are valid without considering the appropriateness of using that model for the data input and the output expected.

• It is not possible to include all future events in a model. For example, a change in legislation could invalidate the results of a model, but may be impossible to predict when the model is constructed.

• It may be difficult to interpret some of the outputs of the model. They may only be valid in relative rather than absolute terms, as when, for example, comparing the level of risk of the outputs associated with different inputs.

1.3 Stochastic and deterministic models

If it is desired to represent reality as accurately as possible, the model needs to imitate the random nature of the variables. A stochastic model is one that recognises the random nature of the input components. A model that does not contain any random component is deterministic in nature.

In a deterministic model, the output is determined once the set of fixed inputs and the relationships between them have been defined. By contrast, in a stochastic model the output is random in nature – like the inputs, which are random variables. The output is only a snapshot or an estimate of the characteristics of the model for a given set of inputs. Several independent runs are required for each set of inputs so that statistical theory can be used to help in the study of the implications of the set of inputs.

A deterministic model is really just a special (simplified) case of a stochastic model.

Whether to use a deterministic or a stochastic model depends on whether you are interested in the results of a single 'scenario' or in the distribution of results of possible 'scenarios'. A deterministic model will give one the results of the relevant calculations for a single scenario; a stochastic model gives distributions of the relevant results for a distribution of scenarios.

1.4 Discrete and continuous states and time

The state of a model is the set of variables that describe the system at a particular point in time taking into account the goals of the study.

Discrete states are where the variables exhibit step function changes in time. For example, from a state of alive to dead, or an increase in the number of cars manufactured in a factory. Continuous states are where the variables change continuously with respect to time. For example, real time changes in values of investments.

The decision to use a discrete or continuous state model for a particular system is driven by the objectives of the study, rather than whether or not the system itself is of a discrete or continuous nature.

A model may also consider time in a discrete or a continuous way.

1.5 Suitability of a model

In assessing the suitability of a model for a particular exercise it is important to consider the following:

- The objectives of the modelling exercise.
- The validity of the model for the purpose to which it is to be put.
- The validity of the data to be used.
- The validity of the assumptions.
- The possible errors associated with the model or parameters used not being a perfect representation of the real-world situation being modelled.
- The impact of correlations between the random variables that 'drive' the model.
- The extent of correlations between the various results produced from the model.
- The current relevance of models written and used in the past.
- The credibility of the data input.
- The credibility of the results output.
- The dangers of spurious accuracy.
- The ease with which the model and its results can be communicated.
- Regulatory requirements.

1.6 Short-term and long-term properties of a model

The stability of the relationships incorporated in the model may not be realistic in the longer term. For example, exponential growth can appear linear if surveyed over a short period of time. If changes can be predicted, they can be incorporated in the model, but often it must be accepted that longer term models are suspect.

Models are by definition, simplified versions of the real-world. They may, therefore, ignore 'higher order' relationships which are of little importance in the short term, but which may accumulate in the longer term.

1.7 Analysing the output of a model

Statistical sampling techniques are needed to analyse the output of a model, as a simulation is just a computer-aided statistical sampling project. The statistician must exercise great care and judgement at this stage of the modelling process as the observations in the process are correlated with each other and the distributions of the successive observations change over time. Therefore we need to be particularly careful before making any assumptions that rely on independence or identical distributions.

1.8 Sensitivity testing

Where possible, it is important to test the reasonableness of the output from the model against the real-world. To do this, an examination of the sensitivity of the outputs to small changes in the inputs or their statistical distributions should be carried out. The appropriateness of the model should then be reviewed, particularly if small changes in inputs or their statistical distributions give rise to large changes in the outputs. In this way, the key inputs and relationships to which particular attention should be given in designing and using the model can be determined.

1.9 Communication of the results

The final step in the modelling process is the communication and documentation of the results and the model itself to others. The communication must be such that it takes account of the knowledge of the target audience and their viewpoint. A key issue here is to make sure that the audience accepts the model as being valid and a useful tool in decision making. It is important to ensure that any limitations on the use and validity of the model are fully appreciated.

2. The Simple Linear Regression Model

2.1. The Model

Let us begin with a simple situation where we have

- one response variable, *Y*
- one explanatory variable, *X*

In many situations,

- *X* can be controlled and is known
- *Y* is unknown but can be observed
- we have *n* pairs of observations { (x_1,y_1), (x_2,y_2), …, (x_n,y_n) }

We would like to use these observations to estimate (or predict) the mean value of *Y* for some given values of *X*. A good place to start exploring the relationship between *X* and *Y* is often to plot them using the *n* pairs of observations. This plot might begin to show the nature of the relationship between *X* and *Y.*

It is good practice to seek the simplest model that describes the relationship well. This idea is called the *principle of parsimony*. At this stage we have not defined what "describes well" means and we will return to this issue a number of times through the module. A *linear* relationship (which would be indicated by something close to a straight-line plot) is the obvious place to start when looking for a simple model.

Given observation data (x_i, y_i) for i = 1, 2, … n we can fit a straight line to describe the response variable *Y* in terms of the explanatory variable *X* where

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

where,

- β_0 denotes the intercept
- β_1 is the slope of the line

However this is a *deterministic* model, meaning it does not allow for any randomness. As such it is unlikely that this model properly describes the data which will usually include some random elements.

We introduce randomness by having a *probabilistic* or *stochastic* element to the model where the model for *Y* has two parts:

- the value for Y we expect to observe for a given value of *X*
- an additional uncontrolled random value

This model can be written either as

$$Y_i = E[Y_i|X = x_i] + \varepsilon_i$$

or as

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i$$

for i = 1, 2, … n

Here ε_i is the *random error.*

It is usual to make the following three standard assumptions about the random error:

(1) $E[\varepsilon_i] = 0$ for all i

(2) $var[\varepsilon_i] = \sigma^2$ for all i

(3) $cov[\varepsilon_i, \varepsilon_j] = 0$ for all $i \neq j$

Because ε_i is a random variable, Y_i is also a random variable. We can re-write the three assumptions above in terms of $Y_i|X = x_i$ rather than ε_i

(1) $E[\,Y_i|X = x_i] = \mu_i = \beta_0 + \beta_1 x_i$ for all i

(2) $var[\,Y_i|X = x_i] = \sigma^2$ for all i

(3) $cov[Y_i|X = x_i,\ Y_j|X = x_j] = 0$ for all $i \neq j$

Putting these three assumptions into words we might say that

(1) the dependence of Y on X is linear

(2) the variance of Y at each value of X is constant and does not depend on x_i

(3) Y_i and Y_j are uncorrelated

Rather than keep writing $Y_i|X = x_i$ we often use $y_i = (Y_i|X = x_i)$ and then the simple linear model can be written as

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

and

$$var[y_i] = \sigma^2$$

It is often convenient to make a further assumption, that the conditional distribution of Y_i is Normal. This is the *Normal Simple Linear Regression Model* which can be written in one of three (equivalent) ways:

(A) $y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = \beta_0 + \beta_1 x_i$

(B) $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

(C) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i are iid $\varepsilon_i \sim N(0, \sigma^2)$

It can be convenient to redefine the parameters of the simple linear model into a *centred* form. This expresses the response variable y_i in terms of both the explanatory variable x_i and the mean level of that explanatory variable $\overline{x}$ where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

if we set

$\alpha = \beta_0 + \beta_1 \overline{x}$ and $\beta = \beta_1$

then the centred form of the model is

$$y_i = \alpha + \beta(x_i - \overline{x}) + \varepsilon_i$$

This model is mathematically identical to the previous (non-centred) form of the model but with new parameters (α and β instead of β_0, β_1) which allows a new interpretation:

- the slope β is the same as that in the previous model β_1
- the new intercept α is the mean response at the mean level of the explanatory variable

2.2. Least Squares Estimation

The model parameters (β_0, β_1 in the simple linear regression model above) are unknown. With a data set we can *estimate* these parameters – that is find values for the parameters that best explain the data we have observed. There are various ways in which parameters can be estimated. Here we consider *least squares* estimation. In later statistics modules you will see other methods e.g. *maximum likelihood estimation*.

The least squares estimators of the model parameters β_0 and β_1 are the parameter values that minimise the sum of the squares of the errors $S(\beta_0, \beta_1)$.

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

To find the minimum we need to differentiate $S(\beta_0, \beta_1)$ with respect to both β_0 and β_1 and set each differential to zero, then solve the two simultaneous equations in β_0 and β_1. The values of β_0 and β_1 that satisfy these simultaneous equations are $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

$$\frac{dS}{d\beta_0} = -2\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0 \qquad (A)$$

and

$$\frac{dS}{d\beta_1} = -2\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]x_i = 0 \qquad (B)$$

if we divide by -2 and separate the items in the brackets in (A) and (B) above we get

$$n\widehat{\beta_0} + \widehat{\beta_1}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (C)$$

and

$$\widehat{\beta_0}\sum_{i=1}^{n} x_i + \widehat{\beta_1}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (D)$$

where (C) and (D) are sometimes called the *normal equations*.

If we divide (C) by *n* we have

$$\widehat{\beta_0} = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta_1}\frac{1}{n}\sum_{i=1}^{n} x_i$$

or

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

and from (D) substituting for $\widehat{\beta_0}$ from above and rearranging gives

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

or

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

which can be written in shorthand as

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

Now in calculus, to check that this is indeed a minimum not a maximum for $S(\beta_0, \beta_1)$ we need to find all the second derivatives $\frac{d^2S}{d\beta_0^2}$, $\frac{d^2S}{d\beta_1^2}$, $\frac{dS}{d\beta_0\beta_1}$ and $\frac{dS}{d\beta_1\beta_0}$ to check that all are > 0.

Note that the equations for the least squares estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ above are functions of *Y* as well as of *X.* Now *Y* is a random variable and is generally unknown. This means that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are also random variables. Because the response variable *Y* is not known, all that we can do is calculate values for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ given a particular set of observations for (x_i, y_i). These values for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are called *least squares estimates.* The *estimator* is the algebraic form depending on the variables X_i and Y_i whilst the *estimate* is that form evaluated for a certain set of observations (x_i, y_i). If we use a different set of observations, we should expect to get different values for the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

3. Assessing the Simple Linear Regression Model

3.1. Properties of the estimators

There are a number of properties of estimators that are desirable. One is for an estimator to be *unbiased*.

If $\hat{\theta}$ is an estimator of θ then we say that $\hat{\theta}$ is an unbiased estimator of θ if $E[\hat{\theta}] = \theta$.

So what about $\widehat{\beta_0}$ and $\widehat{\beta_1}$ in our Normal Simple Linear Regression Model. Are they unbiased?

We will begin with the estimator of the slope parameter, $\widehat{\beta_1}$

Recall from section 2.2 above that

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

which means that $\widehat{\beta_1}$ can be expressed as a function of Y_i in the form

$$\widehat{\beta_1} = \sum_{i=1}^{n} c_i Y_i$$

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

Now under our Normal Simple Linear Regression Model, we assume that the Y_i are independent and normally distributed,

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i are iid $\varepsilon_i \sim N(0, \sigma^2)$

so

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

We know from MTH5129 Probability & Statistics II that a linear combination of independent normal random variables is itself normally distributed. This means that if Y_i follows a Normal distribution, then $\widehat{\beta_1}$ will follow a Normal distribution as well.

To determine whether $\widehat{\beta_1}$ is an unbiased estimator we need to find E[$\widehat{\beta_1}$]

$$E[\widehat{\beta_1}] = E\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i E[Y_i] = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

but $\sum_{i=1}^{n} c_i = 0$ because $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ from the definition of $\bar{x}$

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

therefore

$E[\widehat{\beta_1}] = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1$ so $\widehat{\beta_1}$ is an unbiased estimator of β_1 □

Now for the variance of $\widehat{\beta_1}$

$$var[\widehat{\beta_1}] = var[\sum_{i=1}^{n} c_i Y_i] = \sum_{i=1}^{n} c_i{}^2 var[Y_i] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2 \sigma^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

so in summary for $\widehat{\beta_1}$, the least squares estimator of the slope parameter in the Normal Simple Linear Regression Model

$$\widehat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

Turning to the intercept parameter β_0

Recall from section 2.2 that

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\,\bar{x}$$

and substituting in our expression for $\widehat{\beta_1}$ in terms of Y_i

$$\widehat{\beta_0} = \bar{Y} - \bar{x}\sum_{i=1}^{n} c_i Y_i = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{x}\sum_{i=1}^{n} c_i Y_i = \sum_{i=1}^{n} Y_i\,(\frac{1}{n} - c_i\bar{x})$$

where c_i is defined as before.

This means that $\widehat{\beta_0}$ can also be expressed as a linear combination of Y_i and therefore by the same reasoning as for $\widehat{\beta_1}$ we find that $\widehat{\beta_0}$ follows a Normal distribution.

then

$$E[\widehat{\beta_0}] = E[\bar{Y} - \widehat{\beta_1}\,\bar{x}] = E[\bar{Y}] - \bar{x}E[\widehat{\beta_1}] = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

so $\widehat{\beta_0}$ is an unbiased estimator of β_0 .

for the variance of $\widehat{\beta_0}$

$$var[\widehat{\beta_0}] = var[\sum_{i=1}^{n} Y_i\,(\frac{1}{n} - c_i\bar{x})] = \sum_{i=1}^{n} \sigma^2\left(\frac{1}{n} - c_i\bar{x}\right)^2 = \sigma^2\sum_{i=1}^{n}(\frac{1}{n^2} - 2\frac{c_i\bar{x}}{n} + c_i^2\bar{x}^2)$$

$$= \sigma^2(\frac{n}{n^2} - 0 + \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2\,\bar{x}^2}{S_{xx}^2} = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$$

putting these together we have, for $\widehat{\beta_0}$, the least squares estimator of the intercept parameter in the Normal Simple Linear Regression Model

$$\widehat{\beta_0} \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$$

3.2. Assessing the model

If our model is

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

then with estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ and a set of observations (x_i, y_i) i=1, 2, …, n we can fit the model and estimate the response variable with

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

where the $\hat{y}_i$ values $\hat{y}_1, \hat{y}_2, \ldots \hat{y}_n$ are the *fitted values* or points on the *fitted regression line* corresponding to the n observed x_i values.

Now the observed values $y_1, y_2, \ldots y_n$ will be different to the fitted values $\hat{y}_1, \hat{y}_2, \ldots \hat{y}_n$ that is the observed values will not all lie on the fitted regression line. We define the *residuals* (sometimes called the *crude residuals*) to be

$$e_i = y_i - \hat{y}_i$$

That is the residuals are the observed values minus the fitted values.

The residuals e_i are estimates of the random errors ε_i in the original model specification.

From the least squares definition of $\hat{\beta}_0$ and $\hat{\beta}_1$ we will see that $\sum_{i=0}^{n} e_i = 0$

$$e_i = y_i - \hat{y}_i = e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})$$

so

$\sum_{i=0}^{n} e_i = \sum_{i=0}^{n} (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) = 0 - 0 = 0$ from the definitions of $\bar{y}$ and $\bar{x}$.

When we found the least squares estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ we used a quantity S which is actually a function of β_0 and β_1 so $S(\beta_0, \beta_1)$ where from section 2.2

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2$$

The value of this function for a given data set (x_i, y_i) evaluated at the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ is called the *Residual Sum of Squares* and is denoted SS_E where

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

For a particular data set, SS_E is the minimum value of $S(\beta_0, \beta_1)$ and is a measure of how well the model fits the data. The SS_E is one of the sources of variance of the y_i around their mean $\bar{y}$.

The total variance of the y_i around their mean $\bar{y}$ can be expressed as the *Total Sum of Squares* denoted SS_T where

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

In the Simple Linear Regression Model we will see that:

Total Sum of Squares = Regression Sum of Squares + Residual Sum of Squares

$SS_T = SS_R + SS_E$

where SS_T and SS_E have already been defined.

This equation is sometimes called the *Analysis of Variance Identity*

The Regression Sum of Squares is $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ which is sometimes referred to as the *Model Fit Sum of Squares*

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

$$= \sum_{i=1}^{n}[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 - 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})]$$

$$= SS_E + SS_R + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

now the third term in this equation becomes, after multiplying out the second bracket,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - \bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}e_i\hat{y}_i - \bar{y}\sum_{i=1}^{n}e_i = \sum_{i=1}^{n}e_i\hat{y}_i - 0$$

$$\sum_{i=1}^{n}e_i\hat{y}_i = \sum_{i=1}^{n}e_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0\sum_{i=1}^{n}e_i + \hat{\beta}_1\sum_{i=1}^{n}e_i x_i = 0 + 0 = 0$$

therefore $SS_T = SS_R + SS_E$ □

That is Total Sum of Squares is made up of:

- the Regression Sum of Squares – the variability in the y_i around their mean $\bar{y}$ which is accounted for by the fitted model, and
- the Residual Sum of Squares - the variability in the y_i accounted for by the difference between observed and fitted values.

This view of the variability in the y_i is often represented in an *Analysis of Variance Table* often called an *ANOVA Table* for short.

3.3 The ANOVA Table

The Analysis of Variance (ANOVA) table is shown below:

Source of variation	d.f.	SS	MS	VR
Regression	$v_R = 1$	SS_R	$MS_R = \frac{SS_R}{v_R}$	$F = \frac{MS_R}{MS_E}$
Residual	$v_E = \text{n} - 2$	SS_E	$MS_E = \frac{SS_E}{v_E}$	
Total	$v_T = \text{n} - 1$	SS_T		

In the ANOVA table, the variability in the y_i is accounted for in four different quantities, each represented by a column in the table:

- degrees of freedom (d.f.)
- Sum of Squares (SS)
- Mean Squares (MS)
- Variance Ratio (VR)

We have already covered Sum of Squares above but will now look at the other quantities in the table.

Degrees of Freedom

If we have *n* observations $y_1, y_2, \ldots, y_n$ and then fix either the sum of them or their mean, we can let the values of y_1 vary and still get that sum or mean, we can let the values of y_1 and y_2 vary and still get that sum or mean, … indeed we can let the values of $y_1, y_2, \ldots, y_{n-1}$ vary, but then we will need a certain value for y_n to get the required sum or mean. So here if we have *n* observations, *n-1* are free to vary but one will need to depend on the others. One way of thinking about this is with *n* observations and a fixed sum or mean, *n-1* are independent and free to vary and 1 is taken up by the fixed sum or mean. An estimate of a parameter will be based on observations or pieces of information. The number of independent observations that are used in the estimation of a parameter are the *degrees of freedom* (often abbreviated d.f.).

With the Total Sum of Squares $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$ we have *n* observations, and one degree of freedom is taken up by the calculation of $\bar{y}$, so SS_T has *n – 1* degrees of freedom in the ANOVA table.

With the Residual Sum of Squares $SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ one degree of freedom is taken up with the estimation of $\hat{\beta}_0$ and one d.f. is taken up with the estimation of $\hat{\beta}_1$, so SS_E has *n – 2* degrees of freedom in the ANOVA table.

As $SS_R = SS_T - SS_E$ we can find the degrees of freedom for the Regression Sum of Squares SS_R by the difference in the d.f. for the Total and Residual Sums of Squares = *(n – 1) – (n – 2)* = 1.

Mean Squares

The MS_R and MS_E in the ANOVA table are a measure of the average variation by Regression and Residuals found by dividing the appropriate Sum of Squares by its degrees of freedom.

Variance Ratio

This ratio measures the variation explained by the model fit relative to that explained by the residuals and is denoted *F*.

$$F = \frac{MS_R}{MS_E}$$

We know from MTH5129 Probability & Statistics II that if random variable *X* follows a Chi-squared distribution on v_1 degrees of freedom and variable *Y* follows a Chi-squared distribution on v_2 degrees of freedom, then $\frac{X/v_1}{Y/v_2}$ follows a *Fisher's F Distribution* often simply called an *F-Distribution* with $v_1$and v_2 degrees of freedom.

This is written as $\mathcal{F}_{v_1,v_2}$ or $\mathcal{F}_{v_2}^{v_1}$ or as $\mathcal{F}(v_1, v_2)$. The F-Distribution is skewed and depends on two parameters (v_1, v_2).

This distribution and the Variance Ratio are particularly useful in the Linear Regression model for testing whether $\beta_1$is statistically different from zero. If $\beta_1 = 0$ then we could replace the full linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with a simpler constant model, $y_i = \beta_0 + \varepsilon_i$.

We will see later in this course that if $\beta_1 = 0$ then the Variance Ratio,

$$F = \frac{MS_R}{MS_E} \sim \mathcal{F}_{n-2}^1$$

So to test the null hypothesis H_0: $\beta_1 = 0$ versus the alternative H_1: $\beta_1 \neq 0$ we use the Variance Ratio, F as a test statistic. We reject H_0 at significance level α if

$$F > \mathcal{F}_{n-2}^1(\alpha)$$

where $\mathcal{F}_{n-2}^1(\alpha)$ is the value such that $P\left(F > \mathcal{F}_{n-2}^1(\alpha)\right) = \alpha$

The ANOVA table can also be used to estimate the variance of the residuals σ^2 (which in the Normal Simple Regression Model is also the variance of the y_i).

The Sums of Squares are all functions of the y_i which means that because the y_i are random variables, the different Sums of Squares are random variables as well. It can be helpful to explore the stochastic properties of the Sums of Squares: their expectation, variance and distribution. We will do this in full later on in the course. For now, we will note without proof that in the simple linear regression model, the expected value of the Residual Sum of Squares is given by

$$E(SS_E) = (n-2)\sigma^2$$

Now

$$MS_E = \frac{SS_E}{v_E} = \frac{SS_E}{n-2}$$

which means that

$$E(MS_E) = \sigma^2$$

so MS_E is an unbiased estimator for σ^2 and is often denoted S^2. This is interesting because MS_E itself is <u>not</u> the sample variance in the full linear regression model.

The final quantity to mention here is the *Coefficient of Determination* denoted R^2 which is usually expressed as a percentage and is the percentage of total variation in the y_i explained by the model fitted. That is

$$R^2 = \frac{SS_R}{SS_T}100\% = \left(1 - \frac{SS_E}{SS_T}\right)100\%$$

where, $R^2 = 0$ means that none of the variability in the data is explained by the regression model, and $R^2 = 100$ means that all the observations fit precisely on the fitted regression line.

Note that R^2 is not an indicator of whether there is a relationship between *Y* and *X* but rather the extent to which that relationship is linear.

3.4 Fitted values and residuals

From section 3.2 above, the *residuals* or *crude residuals* are e_i where

$$e_i = y_i - \hat{y}_i$$

which we can also write as

$$e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

or as

$$e_i = y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})$$

and that $\sum_{i=1}^{n} e_i = 0$.

Now $E(e_i) = E\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = E(y_i) - E\left(\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = (\beta_0 + \beta_1 x_i) - (\beta_0 + \beta_1 x_i) = 0$

So the mean of the i[th] residual is zero.

The variance of e_i is given by

$$var(e_i) = \sigma^2(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}})$$

We will not derive this (or the covariance term below) from first principles in this module.

Note though that $var(e_i)$ is not the same as $var(\varepsilon_i)$ which is a constant, σ^2 whereas the expression for $var(e_i)$ includes x_i so it is different for each *i*.

The covariance of two residuals e_i and e_j is given by

$$cov(e_i, e_j) = -\sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}})$$

which again is different from $cov(\varepsilon_i, \varepsilon_j) = 0$.

Therefore from the variance and covariance terms we see that the residuals of the fitted model (e_i) do not behave in exactly the same way as the error term in the original model specification (ε_i).

Therefore rather than crude residuals (e_i) it is sometimes useful to consider *standardised residuals* sometimes denoted d_i. The standardised residuals are designed to have a variance that is closer to the constant σ^2 and covariances that are closer to zero.

$$d_i = \frac{e_i}{[s^2(1 - v_i)]^{\frac{1}{2}}}$$

where,

$$v_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Residual Plots can be a useful way of checking a linear regression model:

- plot the d_i against the x_i to check whether a linear model is appropriate and to see whether the Normal assumptions are appropriate
- plot the d_i against the fitted $\hat{y}_i$ to check for a constant variance (which is called *homoscedasticity*)

To check the assumption of normality (that the errors follow a Normal distribution) we can also use a QQ Plot. If the residual data is from a Normal distribution, then the QQ Plot will be close to a straight line. Points on the QQ Plot away from a straight line suggest that the residuals follow some other, non-Normal, distribution. The QQ Plot is a good first indication but later in the module we will look at a more formal statistical test of the hypothesis that the errors are normally distributed.

4 Inference about the regression parameters

In our simple linear regression model of

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i are iid $\varepsilon_i \sim N(0, \sigma^2)$

we have found (see section 2.2 above) that the least squares estimates of β_0 and β_1 are given by

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

and

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

4.1 Confidence Interval for β_1

We found earlier that the sampling distribution of $\widehat{\beta_1}$ is

$$\widehat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

(Note that even where the y_i are not normally distributed the distribution of $\widehat{\beta_1}$ is approximately normal.)

We can standardise the $\widehat{\beta_1}$, that is find the function of $\widehat{\beta_1}$ that follows a standard normal N(0,1) distribution,

$$\frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

However σ^2 is generally not known, so we will need to replace it with it's estimate s^2. When we do this, the normal distribution becomes a Student t-distribution.

That is because, more generally, if $Z \sim$ N(0,1) and $U \sim \chi^2_\upsilon$ then $\frac{Z}{\sqrt{U}/\upsilon} \sim t_\upsilon$

We already have

$$Z = \frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

and we will see later in the course that

$$U = \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

therefore $T = \frac{\frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2(n-2)}}} = \frac{\widehat{\beta_1} - \beta_1}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$

If we have some parameter Θ a 95% confidence interval for Θ means to find boundaries *a* and *b* such that $P(a < \theta < b) = 0.95$. More generally a 100$(1 - \alpha)$% confidence interval for Θ is to find *a* and *b* such that $P(a < \theta < b) = 1 - \alpha$.

In practice, a confidence interval for β_1 will depend on the data and the estimate $\widehat{\beta_1}$ found from that data. Using the Student-t distribution above, and defining $t_{\frac{\alpha}{2}}$ to be the quantity such that

$$P\left(|t_v| < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

then

$$P\left(\widehat{\beta_1} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}} < \beta_1 < \widehat{\beta_1} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}}\right) = 1 - \alpha$$

So for a particular data set where $\widehat{\beta_1}$ and *S* become values from observed data rather than random variables, we can calculate the 100$(1 - \alpha)$% confidence interval [*a*, *b*] for β_1 where

$$[a, b] = \left[\widehat{\beta_1} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}}, \quad \widehat{\beta_1} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{xx}}}\right]$$

4.2 Testing the significance of β_1

In section 3.3 above we saw that we can test the null hypothesis H_0: $\beta_1 = 0$ using the ANOVA table and the *F* statistic. There is another way to test the same null hypothesis based upon how we have derived the confidence interval for β_1.

Under this null hypothesis, the slope is zero and therefore we have a constant model that can be written $y_i = \beta_0 + \varepsilon_i$

We can test this null hypothesis using the test statistic *T* developed above for confidence intervals. If H_0 is true then β_1 is zero and so

$$T = \frac{\widehat{\beta_1}}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

For a given data set we can calculate the value of *T*. We then reject the null hypothesis H_0: $\beta_1 = 0$ at significance level α if

$$|T| > t_{n-2,\frac{\alpha}{2}}$$

This methodology is in fact equivalent mathematically to the ANOVA table F-statistic route.

Sometimes you will see equations such as those above for the confidence interval and the test statistic *T* written in terms of the estimated *standard error* of $\widehat{\beta_1}$. The standard error $se(\widehat{\beta_1})$ is the square root of the variance of $\widehat{\beta_1}$. Our estimate of the standard error of $\widehat{\beta_1}$ is

$$\widehat{se(\beta_1)} = \sqrt{\frac{S^2}{S_{xx}}}$$

and using the standard error notation, the confidence interval becomes

$$[a,b] = \left[\widehat{\beta_1} - t_{\frac{\alpha}{2}}\, \widehat{se(\beta_1)}\,, \qquad \widehat{\beta_1} + t_{\frac{\alpha}{2}}\, \widehat{se(\beta_1)}\right]$$

and the *T* test statistic is

$$T = \frac{\widehat{\beta_1}}{\widehat{se(\beta_1)}} \sim t_{n-2}$$

4.3 Inference about β_0

Because in modelling we are generally interested in the relationship between *Y* and *X*, we are usually most interested in parameter β_1. We can however also develop confidence intervals and test hypotheses for β_0. We found earlier that the sampling distribution of β_0 is,

$$\widehat{\beta_0} \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$$

Using the same methodology as for $\widehat{\beta_1}$ above, we find that the 100$(1-\alpha)$% confidence interval for β_0 is

$$[a,b] = \left[\widehat{\beta_0} - t_{\frac{\alpha}{2}}\, \widehat{se(\beta_0)}\,, \qquad \widehat{\beta_0} + t_{\frac{\alpha}{2}}\, \widehat{se(\beta_0)}\right]$$

and the test statistic to test the null hypothesis H_0: $\beta_0 = B$ for some value B (which may or may not be zero) is

$$T = \frac{\widehat{\beta_0} - B}{\widehat{se(\beta_0)}} \sim t_{n-2}$$

where $\widehat{se(\beta_0)} = \sqrt{S^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$

4.4 Inference about the mean response

We may also develop confidence intervals and test hypotheses for the mean of the response variable given some value of the explanatory variable, that is $E[Y_i|X_i = x_i]$ which is also often written as μ_i.

Under the simple linear regression model,

$$\mu_i = E[Y_i|X_i = x_i] = \beta_0 + \beta_1 x_i$$

and the least squares estimator of μ_i is given by

$$\widehat{\mu_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

so that for any value of the explanatory variable x_i we can estimate the mean response.

Under the simple linear regression model, the sampling distribution of μ_i is also Normal,

$$\widehat{\mu_i} \sim N(\mu_i,\ \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}))$$

which allows us to obtain a 100$(1-\alpha)$% confidence interval for $\widehat{\mu_i}$ which is

$$[a, b] = \left[\widehat{\mu_i} - t_{\frac{\alpha}{2}}\, \widehat{se(\mu_i)}\,,\qquad \widehat{\mu_i} + t_{\frac{\alpha}{2}}\, \widehat{se(\mu_i)}\right]$$

and we can test the null hypothesis, H_0: $\mu_i = M$ for some value M, with the test statistic

$$T = \frac{\widehat{\mu_i} - M}{\widehat{se(\mu_i)}} \sim t_{n-2}$$

where $\widehat{se(\mu_i)} = \sqrt{S^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}})}$

We should note here though that the value of x_i should be within the range of observed data values for *X* for this estimation of the mean response to be valid. The model has said nothing about the applicability of the linear regression beyond that data range and this should not be used as a method of extrapolation. What we can do though, and will consider next, is to use the model to predict the value of the response variable when presented with some new value for x_i for which y_i has not yet been observed.

4.5 A Prediction Interval for a new observation

More precisely we can develop what is known as a Prediction Interval (sometimes just PI) for some new observation. Let us say that we have a new value for x_i which we will label x_0. We have yet to observe the response for x_0 but we wish to predict it, which we will do by way of an interval rather than a single value given the stochastic nature of our model.

We seek y_0 where $y_0 = \mu_0 + \varepsilon_0$ and the point "prediction" for this would be

$$\widehat{y_0} = \widehat{\mu_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

We know that

$$\widehat{\mu_0} \sim N(\mu_0,\ \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

and therefore the distribution of $\widehat{\mu_0} - \mu_0$ is

$$\widehat{\mu_0} - \mu_0 \sim N(0,\ \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

To gain a prediction interval we would like to have the distribution for $\widehat{y_0} - y_0$ rather than $\widehat{\mu_0} - \mu_0$

So taking our previous equation and then adding and subtracting ε_0 to the left-hand side

$$\widehat{\mu_0} - (\mu_0 + \varepsilon_0) + \varepsilon_0 \sim N(0,\ \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

but the term in the brackets $(\mu_0 + \varepsilon_0)$ is y_0 and $\widehat{y_0} = \widehat{\mu_0}$ so we can re-write this equation as

$$\widehat{y_0} - y_0 + \varepsilon_0 \sim N(0, \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$$

and because we know $\varepsilon_0 \sim N(0, \sigma^2)$ from the original specification of the simple linear model we have

$$\widehat{y_0} - y_0 \sim N(0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) + \sigma^2)$$

or

$$\widehat{y_0} - y_0 \sim N(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right))$$

To find a formula for the prediction interval we need to standardise the normal distribution, that is find the function of $\widehat{y_0} - y_0$ that follows N(0,1).

$$\frac{\widehat{y_0} - y_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{x_0 - \bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

and if we replace σ^2 with its estimator S^2 we have

$$\frac{\widehat{y_0} - y_0}{\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

which allows us to find the 100$(1-\alpha)$% prediction interval for y_0 which is

$$\widehat{y_0} \pm t_{\frac{\alpha}{2}}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

The prediction interval for y_0 is usually much wider than the confidence interval for μ_0 at the same value of α. This is because the random variability term ε_0 impacts the prediction interval.

5 Further Model checking

5.1 Outliers

In regression, an outlier is a single observation where the absolute value of the standardised residual is large compared to the rest of the observations. Outliers are usually obvious in residual plots such as QQ plots.

The standardised residual was defined in section 3.4 as

$$d_i = \frac{e_i}{[s^2(1 - v_i)]^{\frac{1}{2}}}$$

or

$$d_i = \frac{y_i - \hat{y}_i}{s\sqrt{(1 - v_i)}}$$

where,

$$v_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

In some literature you will find suggestions for simple rules for what size of standardised residual constitutes an outlier (e.g. some people suggest $|d_i|$>2). However, what residual values constitute an outlier should depend on the sample size *n.* If we take a statistical approach and calculate what maximum $|d_i|$ would represent a critical value in a test of significance at 95% we get the following:

Sample size *n*	**maximum $|d_i|$ at 95% significance**
6	1.93
8	2.20
10	2.37
20	2.77
30	3.06
60	3.23

If we discover an outlier, the first step is to check the data for any mistakes. If the data does not appear to be an error, then the next step is to re-run the regression analysis with the outlier excluded. If the model results differ from the original, then both should be presented.

5.2 Leverage

Outliers are where one y_i is different from the others. We can also have cases where one x_i is different. This is more of a problem with multiple regression models which we consider later in the course, but we will look at the detection of unusual x_i now in the context of the simple linear regression model. We use *leverage* or v_i which was part of the calculation of standardised residuals in section 3.4 but not discussed further at that time.

$$v_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Now $\sum_i v_i = 2$ so with *n* observations, each on average will have leverage of $\frac{2}{n}$. We generally consider an observation with $v_i > \frac{4}{n}$ as having large leverage. If $v_i > \frac{6}{n}$ then leverage is very high and it is best to check the data for any errors in the recording of the relevant x_i value. Large leverage means that the observation is *influential* and taking that observation out would cause a large change in the β parameter estimates.

We can measure the amount of influence any one observation has using *Cook's Statistic* often labelled D_i. We first perform a simple linear regression on *n* (x,y) observations and find $\widehat{\beta_0}, \widehat{\beta_1}$ and hence $\hat{y}$ values. Then if we omit the observation (x_i , y_i) and repeat the linear regression to gain new parameters and new fitted values denoted $\hat{y}^{(i)}$, Cook's Statistic *for case i* is

$$D_i = \frac{1}{2S^2} \sum_{j=1}^{n} (\hat{y}_j^{(i)} - \hat{y}_j)^2$$

It can be shown that

$$D_i = \frac{1}{2} d_i^2 \frac{v_i}{1 - v_i}$$

This second formula for D_i shows that that Cook's Statistic depends on both the standardised residual d_i and the leverage v_i.

One way to use this statistic to see whether an observation is influential is to compare the D_i value for that observation with the 50th percentile of the F_{n-2}^2 distribution. Another way is to rank all of the D_i values and any that are noticeably larger than the others would suggest an influential observation.

Influential observations do not need to be removed in the way that outliers do but any conclusions from a modelling exercise should note that the results would be different without the influential observation.

5.3 Transformation of the Response

If upon checking the model results, we find that the variance is not constant or that the data is not from a Normal distribution, it might be possible to obtain a better model by some simple transformation of the y_i.

If the data is all non-negative, then the most usual transformation to try first is $\ln y$.

Commonly used transformations and the conditions under which they work best are:

$\ln y$	where Var(Y) is proportional to $E(Y)^2$
$\sqrt{y}$	where Var(Y) is proportional to E(Y), often useful when the data is a count
$sin^{-1}(\sqrt{y})$	often useful if the data is proportions
$1/y$	

5.4 Pure Error and Lack of Fit

If our analysis of the residuals suggests that the data is not from a Normal distribution with a constant variance (the underlying assumption of the simple linear regression model) this means that a straight line regression is not a good model choice. We can generally see this from residual plots, but here we show how to test for this lack of fit more formally.

One possible reason for this which we have not explored so far is *replications*, that is where there are multiple different *y* observations that have the same x_i value.

For notation we use y_{ij} to be the j^{th} observation at x_i where *i* = 1, 2, … m and *j* = 1, 2, … n_i

In the simple linear regression model, although each of the y_{ij} observations might well be different at a certain x_i, the fitted value will be the same $\hat{y}_i$ for all j.

The residuals are now

$$e_{ij} = y_{ij} - \hat{y}_i$$

But now the differences between observed and fitted values come from two sources:

- random variation in y_{ij} where observations at the same x_i can produce different *y* values
- lack of fit in the model which does not capture all that is found in the observed data

We can distinguish between these two sources of residual error.

The *pure error* measures the amount of random variation at x_i and is the difference between an observation y_{ij} and the mean of observations taken at the same x_i.

Pure Error = $y_{ij} - \bar{y}_i$

The *lack of fit* is the difference between the mean observed value and the model fitted value at x_i.

Lack of Fit = $\bar{y}_i - \hat{y}_i$

And so Residual Error = Pure Error + Lack of Fit

More generally we can split the residual sum of squares SS_E into a *pure error sum of squares* SS_{PE} that measures overall random variation, and a *lack of fit sum of squares* SS_{LoF} that measures overall model lack of fit.

Using the *ij* notation we have

$$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2$$

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

$$SS_{LoF} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

and in the simple linear regression model we have

$$SS_E = SS_{PE} + SS_{LoF}$$

Using this we can expand the ANOVA table where there are replications (multiple different y_i observations at the same x_i) splitting SS_E into pure error and lack of fit.

We first need to apportion the n – 2 residual degrees of freedom between *PE* and *LoF*. To calculate SS_{PE} we need to find *m* sample means, the $\bar{y}_i$ for *i* = 1,2,… m and each of these calculations takes up a degree of freedom. Therefore the degrees of freedom for Pure Error are $n - m$.

This leaves $(n-2) - (n-m) = m - 2$ degrees of freedom for Lack of Fit.

For the Mean Squares (MS) column of the ANOVA table we will see later in the course that

$E[SS_{PE}] = (n-m)\sigma^2$ whether the model is true or not, and that

$E[SS_{LoF}] = (m-2)\sigma^2$ if the model is true.

Therefore MS_{PE} gives an unbiased estimator of σ^2 and furthermore MS_{LoF} can give an unbiased estimator of σ^2 if the regression model is true.

Thus in all circumstances,

$$\frac{(n-m)MS_{PE}}{\sigma^2} \sim \chi^2_{n-m}$$

and if the regression model is true,

$$\frac{(m-2)MS_{LoF}}{\sigma^2} \sim \chi^2_{m-2}$$

So finally, for the Variance Ratio (VR) column of the ANOVA table, if the regression model is true then the ratio of the two chi-squared statistics above, each divided by their respective degrees of freedom, follows a F^{m-2}_{n-m} distribution,

$$\frac{MS_{LoF}}{MS_{PE}} \sim F^{m-2}_{n-m}$$

We can now set out the expanded ANOVA table for the case where there are replications in the observations, and we are able to split residual error between pure error and lack of fit.

Source of variation	d.f.	SS	MS	VR
Regression	1	SS_R	MS_R	$\frac{MS_R}{MS_E}$
Residual	n – 2	SS_E	$MS_E = \frac{SS_E}{n-2}$	
Lack of Fit	$m-2$	SS_{LoF}	$MS_{LoF} = \frac{SS_{LoF}}{m-2}$	$\frac{MS_{LoF}}{MS_{PE}}$
Pure Error	$n-m$	SS_{PE}	$MS_E = \frac{SS_{PE}}{n-m}$	
Total	n – 1	SS_T		

We now have a lot of information to take into account when assessing a model:

- residual plots
- ANOVA table
- significance tests on individual parameters
- outliers
- influential observations

6 Matrix approach to Simple Linear Regression

6.1 Re-writing the model in matrix form

Simple linear regression models can also be fitted using matrix approaches. We can think of the previous simple linear regression model based on *n* observations for (x_i, y_i) as a set of *n* equations:

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$

…

$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

Now these same *n* equations can be re-written using matrices and vectors.

If,

- $\boldsymbol{Y}$ is a (nx1) vector of observations y_i
- $\boldsymbol{X}$ is a (nx2) matrix called the *design matrix* where the first column is a series of 1 and the second column is the set of observations x_i
- $\boldsymbol{\beta}$ is a (2x1) vector of the unknown parameters β_0 and β_1

then the *n* equations can be rewritten

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

This way of writing the simple linear model is sometimes called the *General Linear Model* (but care is needed here not to confuse that terminology with *Generalised Linear Modelling* or GLM which is a different form of statistical modelling you will encounter in later statistics modules).

Now $\boldsymbol{Y}$ and ε here are random vectors, that is they are vectors whose elements are random variables. Before we can fit the simple linear regression model in matrix form we need to cover some properties of random vectors and also introduce the Multivariate Normal Distribution as a more general case of the usual Normal Distribution used so far.

6.2 Random Vectors

The first property of random vectors we will need is that the expected value of a random vector is the vector of expected values of the components of that random vector.

So if $z = (z_1, \ldots, z_n)^T$ is a random vector then

$$E[z] = E\begin{pmatrix} z_1 \\ z_2 \\ \cdots \\ z_n \end{pmatrix} = \begin{pmatrix} E[z_1] \\ E[z_2] \\ \cdots \\ E[z_n] \end{pmatrix}$$

We also have properties for expectation of linear transformations of random vectors which are analogous to the properties for single random variables. So if *a* is a constant, *b* is a constant vector, and $\boldsymbol{A}$, $\boldsymbol{B}$ are matrices of constants, then

- $E[az + b] = aE[z] + b$
- $E[\boldsymbol{A}z] = \boldsymbol{A}E[z]$
- $E[z^T\boldsymbol{B}] = E[z]^T\boldsymbol{B}$

With random vectors, variances and covariances of the random variables z_i together form the *dispersion matrix* sometimes called the *variance-covariance matrix.*

$$Var(z) = \begin{pmatrix} var(z_1) & \cdots & cov(z_1, z_n) \\ \vdots & \ddots & \vdots \\ cov(z_n, z_1) & \cdots & var(z_n) \end{pmatrix}$$

- $Var(z)$ can also be expressed as $E[(z - E[z])(z - E[z])^T]$
- the dispersion matrix is symmetric since $cov(z_i, z_j)$ = $cov(z_j, z_i)$
- if all of the z_i are uncorrelated all $cov(z_i, z_j)$ = 0 and hence the dispersion matrix is diagonal with the variances
- if $\boldsymbol{A}$ is a matrix of constants then $Var(\boldsymbol{A}z) = \boldsymbol{A}\,Var(z)\,\boldsymbol{A}^T$

6.3 The Multivariate Normal Distribution

MTH5129 Probability & Statistics II introduced the Bivariate Normal Distribution. We will now extend this to a general case where there are more than two random variables, known as the Multivariate Normal Distribution.

A random vector *z* has a multivariate normal distribution if its probability density function (pdf) can be written in the form

$$f(z) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{V})}} \exp\{-\frac{1}{2}(z - \mu)^T\boldsymbol{V}^{-1}(z - \mu)\}$$

where,

- vector μ is the mean of *z*
- $\boldsymbol{V}$ is the dispersion matrix of *z*
- $\det(\boldsymbol{V})$ is the determinant of $\boldsymbol{V}$

With the multivariate normal distribution we typically use the notation $z \sim N_n(\mu, \boldsymbol{V})$

6.4 Least Squares Estimation using matrices

We are now ready to consider least squares estimation in the general linear model using matrices. Our goal is to find $\widehat{\boldsymbol{\beta}}$ a (2x1) vector with the least squares estimates of the model parameters β_0 and β_1.

When we estimated parameters β_0 and β_1 in the simple linear regression model before we solved the two simultaneous "normal equations" found from taking the derivative of the equation for the sum of squares of errors with respect to each of the two parameters. In matrix form and with our general linear model above, the normal equations become,

$$\boldsymbol{X^T y} = \boldsymbol{X^T X}\,\widehat{\boldsymbol{\beta}}$$

Now as long as $\boldsymbol{X^T X}$ is invertible, that is its determinant is not zero, then there is a unique solution to the matrix form normal equations given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$$

In the simple linear regression model,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

therefore

$$\boldsymbol{X^T y} = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i\, y_i \end{pmatrix}$$

and

$$\boldsymbol{X^T X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

which means that the determinant of $\boldsymbol{X^T X}$ is

$$\left|\boldsymbol{X^T X}\right| = n\sum x_i^2 - \left(\sum x_i\right)^2 = n\, S_{xx} \neq 0$$

hence there is a solution to the normal equations.

The inverse of $\boldsymbol{X^T X}$ is given by

$$(\boldsymbol{X^T X})^{-1} = \frac{1}{n\, S_{xx}}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} = \frac{1}{S_{xx}}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

which means we now have all the components we need to solve the normal equations in matrix form.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$$

$$\widehat{\boldsymbol{\beta}} = \frac{1}{S_{xx}}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} \sum y_i \\ \sum x_i\, y_i \end{pmatrix}$$

$$\widehat{\boldsymbol{\beta}} = \frac{1}{S_{xx}}\begin{pmatrix} \frac{1}{n}\sum x_i^2 \sum y_i - \bar{x}\sum x_i\, y_i \\ \sum x_i\, y_i - \bar{x}\sum\ y_i \end{pmatrix} = \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}S_{xx} - \bar{x}S_{xy} \\ S_{xy} \end{pmatrix} = \begin{pmatrix} \bar{y} - \widehat{\beta_1}\bar{x} \\ \widehat{\beta_1} \end{pmatrix}$$

which is identical to the previous result for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ in the simple linear regression model not in matrix form.

Then the fitted values in matrix form are then,

$$\widehat{\boldsymbol{\mu}}_i = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and the Residual Sum of Squares in matrix form is

$$SS_E = \boldsymbol{y}^T \boldsymbol{y} - \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}$$

which if you complete all the matrix multiplication gives

$$SS_E = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

6.5 Properties that follow from the matrix approach

There follows a number of theorem and lemmas that flow from the matrix approach parameters and residuals which we will present here.

(a) The least squares estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ that is $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

(b) $Var[\boldsymbol{\beta}] = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$

(c) If, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ and $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$ then $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1})$

(d) The vector of fitted values, $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ can be written in the form $\widehat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{H}$ is called the *hat matrix* and is given by $\boldsymbol{H} = \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ and $\boldsymbol{H}$ has the two properties that $\boldsymbol{H} = \boldsymbol{H}^T$ and $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ (this second property is called an *indempotent matrix*).

(e) If the residual vector is $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ then $E[\boldsymbol{e}] = \mathbf{0}$

(f) $Var[\boldsymbol{e}] = \sigma^2 (\boldsymbol{I} - \boldsymbol{H})$

(g) The sum of squares of the residuals is $\boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{Y}$

(h) The elements of the residual vector $\boldsymbol{e}$ sum to zero, that is $\sum_{i=1}^{n} e_i = 0$

(i) Because of the result (h) above and all the e_i sum to zero, we also have $\frac{1}{n}\sum \widehat{Y}_i = \bar{Y}$

The centred form of the simple linear regression model can also be written in matrix or general linear form. From before the centred form was $y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i$

Now in matrix form and centred we use

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

and the results which follow are

$$\hat{\alpha} = \bar{y}$$

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

$$var[\hat{\alpha}] = \sigma^2/n$$

$$var[\hat{\beta}] = \frac{\sigma^2}{S_{xx}}$$

and

$$cov[\hat{\alpha}, \hat{\beta}] = 0$$

This last result, that $\hat{\alpha}$ and $\hat{\beta}$ are uncorrelated, can make this centred form useful in certain areas of practical work.

6.6 Maximum Likelihood Estimation

So far, we have used least squares estimation to find our model parameter estimators $\widehat{\boldsymbol{\beta}}$. There are other ways of finding estimates for parameters in a model and we will now consider one here that is widely used beyond the simple linear regression model. This is Maximum Likelihood Estimation (MLE) which you will encounter in a number of different contexts and with various probability distributions, in later statistics modules.

Let us say we have a set of *n* observations Y_1, Y_2, … Y_n which are assumed to be independent observations which all come from the same probability distribution.

Now let us say that the probability distribution from which these are assumed to come has a probability density function $f(y_i)$ which has a parameter θ so that the pdf can be written $f(y_i|\theta)$. The parameter θ is unknown and we wish to estimate it by Maximum Likelihood Estimation.

The maximum likelihood estimator of θ is that value of θ which maximises the joint probability that the *n* observations occur. To find this probability to maximise we develop something called the Likelihood function which is usually written $L(\theta, y)$ or sometimes just $L(\theta)$ and is a function of θ.

$$L(\theta, y) = \prod_{i=1}^{n} f(y_i|\theta)$$

And for discrete observations this becomes

$$L(\theta, y) = \prod_{i=1}^{n} Pr(Y_i = y_i|\theta)$$

The maximum likelihood estimator written $\hat{\theta}$ is that value of θ which maximises the Likelihood function $L(\theta, y)$.

Once again, we will use calculus to find the estimator. In least squares estimation we differentiated the sum of squares equation with respect to the model parameters β_0 and β_1 and set to zero to find a minimum. Here we will differentiate the Likelihood function with respect to the parameters and set to zero to find a maximum.

In most cases of MLE for probability distributions it is easier to take the log of the likelihood function and differentiate $\log\ L(\theta, y)$ rather than $L(\theta, y)$. The $\hat{\theta}$ that maximises $\log\ L(\theta, y)$ will be the same as the one that maximises $L(\theta, y)$.

Before we look at MLE for the Normal distribution and its application to the simple linear regression model, let us look at MLE for a more straightforward probability distribution, the Binomial.

Let us say that we have *n* binomial trials where $y_i = 1$ if the i^{th} trial is a success and $y_i = 0$ otherwise.

Let the probability of a success be *p* (which is unknown and we seek to estimate from the *n* observations). We seek the Maximum Likelihood Estimator of *p* the Binomial success parameter.

If $y = \sum_{i=1}^{n} y_i$ that is the total number of successful trials,

Then the Likelihood function is

$$L(p) = L(y_1 \dots y_n|p) = p^y(1-p)^{n-y}$$

And we seek $\hat{p}$ which is the value of *p* that maximises $L(p)$ by differentiating and setting to zero.

As $L(p)$ is a product of functions, it will be easier to differentiate *log* $L(p)$

$$logL(p) = \log(p^y(1-p)^{n-y}) = y\log(p) + (n-y)\log(1-p)$$

And

$$\frac{dlogL(p)}{dp} = y\frac{1}{p} - (n-y)\frac{1}{1-p}$$

If we set this to zero and solve for *p*

$$y\frac{1}{\hat{p}} - (n-y)\frac{1}{1-\hat{p}} = 0$$

$$\frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} = 0$$

$$y(1-\hat{p}) = (n-y)\hat{p}$$

$$y = n\hat{p}$$

$$\hat{p} = \frac{y}{n}$$

So the MLE for Binomial parameter *p* is the proportion of observed trials that are successful.

To complete this we should take second derivatives to see that we have found a maximum not a minimum for the log likelihood.

The Binomial example highlights one of the key properties of (and advantages of) maximum likelihood estimators. With this Binomial case we would expect the quality of the estimate to increase with sample size *n*. Statistically we say that the estimator has strong *asymptotic* properties, that is as $n \rightarrow \infty$

In particular, maximum likelihood estimators are:

- Asymptotically unbiased
- Normally distributed
- Achieve the smallest variance possible.

But the Binomial example also highlights the key weakness

- At small *n* the estimator can be biased
- In general the asymptotic properties may not apply at smaller sample sizes.

We can now move to MLE in the Normal distribution which we will need to apply maximum likelihood in the simple linear regression model.

For a normal distribution with mean μ and variance σ^2 we can estimate μ by MLE. We begin with the Normal pdf

$$f(y|\mu) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

And so the likelihood function is

$$L(\mu, y) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (y - \mu)^2\right)$$

And taking logs

$$logL = -\log\left(\sigma^n (2\pi)^{\frac{n}{2}}\right) - \frac{1}{2\sigma^2} \sum (y - \mu)^2$$

Differentiating

$$\frac{dlogL}{d\mu} = \frac{1}{\sigma^2} \sum (y - \mu)$$

Which equals zero when $\hat{\mu} = \bar{y}$

Now in our simple linear regression model instead of $Y_i \sim N(\mu, \sigma^2)$ we now have

$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and we seek to estimate β_0 and β_1 by MLE.

Now the likelihood function becomes a function of the two model parameters rather than of the normal mean

$$L(\beta_0, \beta_1, y_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (y_i - \beta_0 + \beta_1 x_i,)^2\right)$$

And the likelihood and the log likelihood are maximised when $-\sum (y_i - \beta_0 + \beta_1 x_i,)^2$ is maximised. Note that this is exactly the same place where $\sum (y_i - \beta_0 + \beta_1 x_i,)^2$ is minimised, which was precisely what we did when we found parameter estimates by least squares.

Therefore in the simple linear regression model, the least squares estimators of β_0 and β_1 are the same as the maximum likelihood estimators.

4 Multiple Linear Regression Model

4.1 Other explanatory variables

Whenever we fit a simple linear regression model there will be some amount of variation in the y_i that is not explained by the regression (that part of the R^2 less than 100%). Part of this remaining variation might be other explanatory variables. A multiple linear regression model is one that seeks to take into account more than one explanatory variable.

If we have 2 explanatory variables X_1 and X_2 and a response variable Y we can write the multiple linear regression model as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

For i = 1, 2, …, n observations of the form (x_{1i} , x_{2i} , y_i)

More generally we can have a multiple linear regression model with p – 1 explanatory variables X_1 , X_2 , … , X_{p-1} which we can write either as

$$E[y_i] = \mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1\,i}$$

$$var(y_i) = \sigma^2 \text{ for all } i = 1, \ldots, n$$

$$cov(y_i, y_j) = 0 \text{ for all } i \neq j$$

Or alternatively and equivalently as,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1\,i} + \varepsilon_i$$

$$var(\varepsilon_i) = \sigma^2 \text{ for all } i = 1, \ldots, n$$

$$cov(\varepsilon_i, \varepsilon_j) = 0 \text{ for all } i \neq j$$

And we usually have the additional assumption of normality which can be written as either $y_i \sim N(\mu_i, \sigma^2)$ or as $\varepsilon_i \sim N(0, \sigma^2)$

We can also write the multiple linear regression model in matrix form. This is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

where,

$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ the vector of responses

$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ the *design matrix*

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ the vector of parameters which are unknowns

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ the vector of random errors

4.2 Least Squares estimation in the multiple regression model

Algebraically we will find it easiest to work with the matrix form to derive the least squares estimates for **β** and then we will find that the results are the same as those found for the simple linear regression model in section 3 above.

Once again to find the least squares estimators we minimise the sum of squares of residuals that is

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1\,i}))^2$$

or

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} {\varepsilon_i}^2$$

$$S(\boldsymbol{\beta}) = \varepsilon^T \varepsilon$$

The least squares estimator $\widehat{\boldsymbol{\beta}}$ of the vector of unknown parameters $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y}$$

This is the same result as in section 3 above except that this time the identity matrix ***X*** has *p* columns for p – 1 explanatory variables whereas before it had 2 columns.

From the work we have already done on the simple linear regression model we also know that:

- $\widehat{\boldsymbol{\beta}}$ is an *unbiased estimator* for $\boldsymbol{\beta}$
- Var[$\widehat{\boldsymbol{\beta}}$] = $\sigma^2 (\boldsymbol{X^T X})^{-1}$
- If $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I})$ then $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X^T X}\right)^{-1})$

In finding the vector of fitted values $\widehat{\boldsymbol{Y}}$ we can use the *hat matrix* ***H*** where

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} = \boldsymbol{HY}$$

So

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1} \boldsymbol{X^T}$$

And recall from section 3 that $\boldsymbol{H^T} = \boldsymbol{H}$ and $\boldsymbol{HH} = \boldsymbol{H}$, the property of an idempotent matrix.

With the hat matrix we can now look at the residual vector ***e***

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{HY} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

Then

$$E[\boldsymbol{e}] = 0$$

Which we can show by:

$$E[\boldsymbol{e}] = (\boldsymbol{I} - \boldsymbol{H})E(\boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T})E[\boldsymbol{Y}] = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T})\boldsymbol{X\beta} = \boldsymbol{X\beta} - \boldsymbol{X\beta}$$

Also

$$var(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

Which we can show by:

$$var(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})var(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{H})^2 = \sigma^2(\boldsymbol{I} - 2\boldsymbol{H} - \boldsymbol{H}^2) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

The sum of all the elements in $\boldsymbol{e}$ is zero which is the same as the $\sum e_i = 0$ result we had before in section 2.

The sum of squares of residuals in matrix form is $\boldsymbol{e^T e}$ and

$$\boldsymbol{e^T e} = \boldsymbol{Y^T}(\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y}$$

4.3 Analysis of Variance

The analysis of variance identity can be used for multiple linear regression and for regression in matrix form in the same way that it was for simple linear regression. That is,

Total sum of squares = Regression sum of squares + Residual sum of squares

$SS_T = SS_R + SS_E$

In matrix form the total sum of squares is

$$SS_T = \sum (Y_i - \bar{Y})^2 = \boldsymbol{Y^T Y} - n\bar{Y}^2$$

And the regression sum of squares is

$$SS_R = \sum (\hat{Y}_i - \bar{Y})^2 = \boldsymbol{Y^T H\, Y} - n\bar{Y}^2$$

We have already seen that the residual sum of squares can be written as

$$SS_E = \sum (Y_i - \hat{Y}_i)^2 = \boldsymbol{Y^T}(\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y}$$

It is possible to combine these to show the analysis of variance identity in matrix form and for multiple linear regression as we previously did with the simple linear regression model.

We can also produce an ANOVA table for a multiple linear regression with *n* observations and *p* – 1 explanatory variables and hence *p* parameters estimated (β_0 , β_1 … β_{p-1})

The ANOVA table is again in the format we have seen before

	d.f.	SS	MS	VR
Regression				
Residuals				
Total				

Where now the Regression row represents the multiple linear regression.

Now the degrees of freedom are:

- $n - p$ for residuals (this is the general case of $n - 2$ when $p = 2$ in the simple linear regression model before)
- $p - 1$ for regression (this is the general case of 1 when $p = 2$ in the simple linear regression model before)
- $n - 1$ in total (as before)

We have already given the formulae for sums of squares. Mean squares are then those sums of squares divided by their respective degrees of freedom.

$$MS_R = \frac{SS_R}{p-1}$$

$$MS_E = \frac{SS_E}{n-p} = S^2$$

And once again $MS_E = S^2$ is an unbiased estimator for σ^2

Then the variance ratio or F statistic becomes

$$VR = \frac{MS_R}{MS_E} = \frac{\frac{SS_R}{p-1}}{S^2}$$

4.4 Overall test of significance of a multiple regression

We can use the Variance Ratio in the multiple regression ANOVA table to test whether the overall multiple regression has significance compared to a "null model" of a constant β_0 plus some random variation ε_i.

Our null hypothesis is

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

And the alternative hypothesis is that at least one of $\beta_1, \beta_2, \ldots \beta_{p-1}$ is not zero.

Our F-statistic, sometimes written *F** is the variance ratio in the ANOVA table

$$F^* = \frac{\frac{SS_R}{p-1}}{\frac{SS_E}{n-p}} = \frac{\frac{SS_R}{p-1}}{S^2}$$

Where the denominator is always an unbiased estimator of σ^2 but the numerator is only an unbiased estimator of σ^2 if the multiple regression assumptions (linear relationships, constant variance and normal distribution) are true.

Under H_0 we will have $F^* \approx 1$ so large values of F^* are required to reject H_0 (which is what we generally seek to do as we would like a model that has significance).

The F-test here compares F^* with the critical value of the Fisher's-F distribution on $p-1$ and $n-p$ degrees of freedom where we reject H_0 at 100(1 – α)% significance if $F^* > F_{n-p}^{p-1}(\alpha)$.

4.5 Inference about parameters in multiple regression models

We already have the distribution of the least squares estimators of the *p* model parameters

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 \, (\boldsymbol{X^T X})^{-1})$$

So if we want the j[th] parameter estimator $\widehat{\beta_j}$ where j = 0, 1, …, $p-1$, then

$\widehat{\beta_j} \sim N(\beta_j \, , \sigma^2 \, c_{jj})$ where c_{jj} is the j[th] diagonal element of $(\boldsymbol{X^T X})^{-1}$ where we count the diagonal elements 0, 1, …, p – 1 (i.e. the first diagonal element relates to β_0, the second one to β_1, and the last one to β_{p-1}.

In this way we can make inference about β_j in the ways in which we did for β_1 in the simple linear regression model earlier. These are:

- Confidence intervals for β_j
- Tests of hypotheses with H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$

In line with the parameter confidence intervals we constructed in the simple linear model, our 100(1 – α)% confidence interval for β_j is

$$[a, b] = \widehat{\beta_j} \pm t_{n-p}(\alpha)\sqrt{S^2 c_{jj}}$$

The test statistic for H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$ is *T* where,

$$T = \frac{\widehat{\beta_j}}{\sqrt{S^2 c_{jj}}} \sim t_{n-p} \text{ under } H_0$$

We need to be very careful about the interpretation of these confidence intervals and tests of hypotheses. They only apply within the context of the whole *p* parameter model that is being fitted.

For example if we cannot reject H_0: $\beta_j = 0$ then:

- This does <u>not</u> mean that X_j has no explanatory power, rather that it has no <u>additional</u> explanatory power compared to the $p - 1$ parameter model that had all of the other betas apart from β_j
- Also this does not tell us about the model $y_i = \beta_0 + \beta_j x_{ji} + \varepsilon_i$ compared to the "null" model $y_i = \beta_0 + \varepsilon_i$, rather it tells us about the role of β_j within the whole p parameter model.

4.6 Confidence Intervals for μ in multiple regression

We might want to estimate the mean response, μ at a certain value of $\boldsymbol{x}$.

We already know that $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{E[Y]}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

Now say we want to estimate μ_0 at $x_0 = (1,\ x_{1,0}\ \ldots\ x_{p-1,0})^T$ where

$\mu_0 = E[Y|X_1 = x_{1,0}\ \ldots\ X_{p-1} = x_{p-1,0}]$

Our point estimate is

$$\hat{\mu}_0 = x_0^T \widehat{\boldsymbol{\beta}}$$

With a multiple linear regression model that includes the assumption of a normal distribution we can develop a confidence interval for μ_0.

Now, $\hat{\mu}_0 = x_0^T \widehat{\boldsymbol{\beta}}$ is a linear combination of the components of $\widehat{\boldsymbol{\beta}}$ all of which are normally distributed therefore $\hat{\mu}_0$ must also be normal.

$E[\hat{\mu}_0] = E\left[x_0^T \widehat{\boldsymbol{\beta}}\right] = x_0^T \boldsymbol{\beta} = \mu_0$ and

$var[\hat{\mu}_0] = var\left[x_0^T \widehat{\boldsymbol{\beta}}\right] = x_0^T var(\widehat{\boldsymbol{\beta}}) x_0 = \sigma^2 x_0^T (\boldsymbol{X^T X})^{-1} x_0$

And putting all these together we have

$\hat{\mu}_0 \sim N(\mu_0, \sigma^2 x_0^T (\boldsymbol{X^T X})^{-1} x_0)$ from which it is straightforward to develop a 100(1 – α)% confidence interval for μ_0 which is

$$[a, b] = \hat{\mu}_0 \pm t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{S^2 x_0^T (\boldsymbol{X^T X})^{-1} x_0}$$

4.7 Prediction Intervals in multiple regression

Now say we have a new set of $\boldsymbol{x}$ observations $x_0 = (1,\ x_{1,0}\ \ldots\ x_{p-1,0})^T$ but we do not yet have the corresponding observation for the response y_0. When we predict y_0 with a prediction interval we will need to take into account the random variation that comes with a new observation.

Our point estimate for y_0 is $\hat{\mu}_0$ which is the same as $\hat{y}_0$

With our Normal distribution assumption for the y_i's we have

$$\hat{y}_0 \sim N(\mu_0, \sigma^2 x_0^T \, (\boldsymbol{X^T X})^{-1} x_0 \,)$$

$$\hat{y}_0 - \mu_0 \sim N(0, \sigma^2 x_0^T \, (\boldsymbol{X^T X})^{-1} x_0 \,)$$

$$\hat{y}_0 - (\mu_0 + \varepsilon_0) \sim N(0, \sigma^2 x_0^T \, (\boldsymbol{X^T X})^{-1} x_0 + \sigma^2)$$

So

$$\hat{y}_0 - y_0 \sim N(0, \sigma^2 (1 + x_0^T \, (\boldsymbol{X^T X})^{-1} x_0))$$

Standardising gives us

$$\frac{\hat{y}_0 - y_0}{\sqrt{\sigma^2 (1 + x_0^T \, (\boldsymbol{X^T X})^{-1} x_0)}} \sim N(0, 1)$$

And replacing the unknown σ^2 with our estimate S^2 gives

$$\frac{\hat{y}_0 - y_0}{\sqrt{S^2 (1 + x_0^T \, (\boldsymbol{X^T X})^{-1} x_0)}} \sim t_{n-p}$$

Which allows us to develop the 100(1 – α)% prediction interval for y_0 which is

$$\hat{y}_0 \pm t_{n-p} \left(\frac{\alpha}{2}\right) \sqrt{S^2 (1 + x_0^T \, (\boldsymbol{X^T X})^{-1} x_0)}$$

5 Model building

In building a multiple regression model we have two objectives which seem to be in conflict:

- having a model that describes the data as well as possible
- having a model that is as simple as possible (the principle of parsimony)

Selecting a model – or a subset of the potential explanatory variables – that gives a suitable balance between these objectives can be more art than science. There is no one correct answer. The interaction between the explanatory variables makes this even more complex because a combination of say three explanatory variables may explain more, or demonstrate better modelling properties (normal distribution, constant variance) than any of the three explanatory variables when used in a simple linear regression.

So in this section we will look at a number of approaches to deciding which explanatory variables to keep in a multiple linear regression model.

5.1 Using the F test to delete variables

Let us say we have a multiple linear regression model with *p – 1* explanatory variables and *p* parameters. With an ANOVA table we can carry out a test of the overall model and see that not all of the β parameters are zero and hence the multiple linear regression model has some significance and some explanatory power. But perhaps we could delete some of the explanatory variables to leave a simpler model that still contains explanatory power.

We do this with a *Subset test*. We are looking to see whether the *p* parameter model could be reduced to a *q* parameter model (*q* < *p*).

We are looking to see whether we can keep $x_1, \dots, x_{q-1}$ but remove $x_q, \dots, x_{p-1}$. *Note that in practice we will not necessarily be keeping variables in number order. For example in a six variable, 7 parameter model where we look to remove 2 variables it is not necessarily the case that* x_5 *and* x_6 *are the variables to be deleted first, but rather the two that contribute least to model significance. We will cover how to identify which variables to consider for deletion later.*

More specifically we are interested in whether these variables under consideration for deletion significantly increase the sum of squares due to regression or significantly reduce the sum of squares due to residuals compared with the simpler model that does not include them. This is sometimes referred to as the "extra sum of squares principle". The idea is that we seek models that maximise the proportion of sums of squares that are due to regression and minimise the proportion due to residuals.

We seek the *extra sum of squares* due to $x_q, \dots, x_{p-1}$ given that $x_1, \dots, x_{q-1}$ are already in the model. This can be written $SS(x_q, \dots, x_{p-1} \mid x_1, \dots, x_{q-1})$

Extra SS = {Regression SS under the full model} – {Regression SS under the reduced model}

and

Extra SS = {Residual SS under the reduced model} – {Residual SS under the full model}

Let $\boldsymbol{\beta_1^T} = (\beta_0, \dots, \beta_{q-1})$ and $\boldsymbol{\beta_2^T} = (\beta_q, \dots, \beta_{p-1})$

so that $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$

that is we have split the parameter vector $\boldsymbol{\beta}$ into one vector for the reduced model with *q* parameters and another vector with the additional *p – q* parameters we are considering for deletion.

similarly we can split up the $\boldsymbol{X}$ matrix into $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ where $\boldsymbol{X_1}$ contains a columns of 1's and then *q – 1* columns with n observations for explanatory variables $x_1, \dots, x_{q-1}$ and $\boldsymbol{X_2}$ contains *p – q* columns with *n* observations for explanatory variables $x_q, \dots, x_{p-1}$.

Then the full model is

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \varepsilon$$

$$\mathbf{Y} = \boldsymbol{X_1}\,\boldsymbol{\beta_1} + \boldsymbol{X_2}\,\boldsymbol{\beta_2} + \varepsilon$$

and the reduced model is

$$\mathbf{Y} = \boldsymbol{X_1}\,\boldsymbol{\beta_1} + \varepsilon$$

We can calculate the SS_R and SS_E for both the full and the reduced models. We will call them:

SS_R^{Full} and SS_E^{Full}

SS_R^{Red} and SS_E^{Red}

these use the same formulae that we developed in the previous section for sums of squares under multiple linear regression models but with the appropriate vector $\boldsymbol{\beta}$ and matrix $\boldsymbol{X}$ for the full / reduced model.

Then extra sum of squares is

$SS_{extra} = SS_R^{Full} - SS_R^{Red} = SS_E^{Red} - SS_E^{Full} = \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{Y} - \widehat{\boldsymbol{\beta}_{\boldsymbol{1}}}^{\boldsymbol{T}}\boldsymbol{X_1}^{\boldsymbol{T}}\,\boldsymbol{Y}$ in matrix form.

Once we have calculated the extra sum of squares we need to test whether that amount is significant or not. We do this with a test of hypotheses.

$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$

H_1: at least one of these parameters is not zero.

If we reject H_0 then there is evidence that at least some of the additional variables $x_q, \dots, x_{p-1}$ are significant and should be included in the model.

If we cannot reject H_0 then we should delete the variables $x_q, \ldots, x_{p-1}$.

Under H_0 the test statistic *F** follows a Fisher F distribution

$$F^* = \frac{\left(\frac{SS_{extra}}{p-q}\right)}{s^2}$$

here s^2 is found from MS_E in the full model

and under H_0 $F^* \sim F_{n-p}^{p-q}$

so we reject H_0 at α significance level if $F^* > F_{n-p}^{p-q}(\alpha)$

We may set out the calculation for this test in a particular form of ANOVA table.

Source	d.f.	SS	MS	VR = F*
$x_1, \ldots, x_{q-1}$ $x_q, \ldots, x_{p-1} \mid x_1, \ldots, x_{q-1}$	q – 1 p – q	$SS(x_1, \ldots, x_{q-1})$ SS_{extra}	$\frac{SS_{extra}}{p-q}$	$\frac{\left(\frac{SS_{extra}}{p-q}\right)}{s^2}$
Overall Regression Residual	p – 1 n – p	SS_R SS_E	s^2	
Total	n – 1	SS_T		

There are two special cases where the F test can be replaced by a t test:

- where p – q = 1 so only one explanatory variable is being considered for deletion
- where there is a natural ordering of the explanatory variables X_1, X_2, X_3, … so that we naturally consider deleting them one at a time sequentially according to that order.

For deleting one explanatory variable (in this case we will consider deleting X_{p-1} but our one variable for deletion does not need to be the one with the highest subscript) our test statistic is

$$t = \frac{\hat{\beta}_{p-1}}{se(\hat{\beta}_{p-1})}$$

where $se(\hat{\beta}_{p-1})$ is the estimated standard error of the relevant beta parameter. The `summary()` function for a `lm()` linear model in R will include this standard error estimate in its output for each coefficient.

Under H_0: $\beta_{p-1} = 0$, this t statistic $t \sim t_{n-p}$ and we complete a two-sided test of *t* at our chosen level of significance. It can be shown that under the null hypothesis $F^* = t^2$.

Where there is a natural ordering of the X_i variables, we can perform a sequence of t tests to consider deletion of these variables in reverse order.

In this case the full model with $p - 1$ explanatory variables whose regression has $p - 1$ d.f. can be thought of as the sum of $p - 1$ consecutive models each of which has one explanatory variable and 1 d.f. These are:

X_1

$X_2 \mid X_1$

$X_3 \mid X_1, X_2$

…

$X_{p-1} \mid X_1, X_2 \ldots X_{p-2}$

We can then test successive β_j parameters starting with $j = p - 1$ and working backwards towards $j = 1$ each with a *t* test using that parameter estimate and its estimated standard error in the presence of prior explanatory variables in the ordering.

5.2 All Subsets Regression

Usually there will not be a natural order to the explanatory variables. In this case there is some advantage to being able to consider all of the possible linear regression models using a set of explanatory variables from the largest (the full model with all $p - 1$ variables and p parameters) down to the constant or null model (with no explanatory variables and 1 parameter β_0 plus random error).

With three explanatory variables the set of all linear regression models is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{3i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_3 x_{3i} + \varepsilon_i$$
$$y_i = \beta_0 + \varepsilon_i$$

More generally with $p - 1$ explanatory variables there will be 2^{p-1} models to consider. This means that even with relatively small p a manual comparison of all models becomes difficult and time consuming.

One method for comparing the full set of models is to calculate one or more statistics for each model and then compare the models using these statistics. The statistics we consider are:

- variance
- $s^2 = MS_E$
- R^2
- Adjusted R^2

- Mallow's Statistic

5.2.1 Variance and MS_E

Having the model with the smallest possible variance of residuals σ^2 is a natural starting point however the true variance is unknown. That leads us quickly to $s^2 = MS_E$ which we know is an unbiased estimator of σ^2. However if our method is simply to select the model that has lowest MS_E, that will often (although not always) be the full model. Therefore this is a very conservative method of model building. Better is to locate the model with the smallest number of explanatory variables that keeps its MS_E close to full model MS_E. Definitions of "close" will vary according to modelling scenarios. A useful way to look at this in practice is to plot all the model MS_E's against number of explanatory variables.

5.2.2 R^2 and Adjusted R^2

From before we have the Coefficient of Determination or R^2 of a model is

$$R^2 = 100\% \frac{SS_R}{SS_T} = 100\%(1 - \frac{SS_E}{SS_T})$$

Adding additional explanatory variables should always increase R^2 therefore once again if our model building rule is simply to maximise R^2 we will always use the full model. Similar to MS_E above a more useful approach is to calculate R^2 for all the models and then plot R^2 against number of explanatory variables to see where the gain in R^2 from adding more variables starts to level off.

R^2 does not in itself take any account of the number of explanatory variables in the model therefore using R^2 to compare say a 2 variable and 5 variable model has disadvantages. Adjusted R^2 (found alongside R^2 in `summary()` output of `lm()` in R) seeks to adjust for this.

$$\text{Adjusted } R^2 = 100\%(1 - (n-1)\frac{MS_E}{SS_T})$$

Adjusted R^2 does take account of the number of explanatory variables. Whereas R^2 always increases when a new variable is added, Adjusted R^2 will only increase if the *F* statistic for the parameter associated with that new explanatory variable is greater than 1. Selecting the model with largest Adjusted R^2 is a method of comparison across models with different number of explanatory variables and seeking to maximise Adjusted R^2 will not automatically lead to the full model.

5.2.3 Mallow's Statistic

For a model with *k* parameters we define Mallow's Statistic, C_k to be

$$C_k = \frac{SS_E^{(k)}}{\sigma^2} + 2k - n$$

where $SS_E^{(k)}$ is the residual sum of squares for the linear regression model with those *k* parameters.

If this *k* parameter model has all the statistically significant explanatory variables available then $E[SS_E^{(k)}] = (n - k)\,\sigma^2$ and then C_k becomes (n – k) + 2k – n = k.

If the model excludes one or more of the statistically significant explanatory variables available then $E[SS_E^{(k)}] > (n - k)\,\sigma^2$ and then $C_k >$ k.

This would suggest choosing the model with Mallow's Statistic C_k closest to its number of parameters k.

However it can also be shown that Mallow's Statistic is also an estimator of the mean square error of prediction in a linear regression model with k parameters. Therefore there is also some utility in minimising C_k. Hence two potential model selection rules using Mallow's Statistic are available:

- select model with C_k closest to *k*
- select model with smallest C_k

In practice σ^2 which is part of the C_k is unknown. We usually replace this with $S^2 = MS_E^{full}$ from the full model (not the k parameter model). R estimates C_k this way. If `full_model` and say `model_k` have both been constructed in R with `lm()` and the appropriate explanatory variables then Mallow's Statistic (sometimes called Mallow's Cp) is found by `ols_mallows_cp(model_k, full_model)`

5.3 Automatic Methods of model selection

If we have a relatively small number of explanatory variables available then calculating MS_E or R^2 or Adjusted R^2 or Mallow's Statistic C_k for each of the candidate models and then making a selection based around some criterion and those statistics is feasible. However, as the number of potential explanatory variables grows (and the number of candidate models grows exponentially) doing this for all models becomes more challenging. In response to this, a number of so-called automatic regression model selection procedures have been devised. Each of these has their advantages and disadvantages. They generally involve a sequence of statistical tests.

5.3.1 Backwards Elimination

The process can be summarised as follows:

- fit the multiple linear regression model that uses all the explanatory variables

- as the number of variables increases and we risk including variables that are essentially themselves linear combinations of other variables in the model, we run into the problem of multicollinearity which we will look at in the section below
- Calculate the *F* statistic (or the *t* statistic) for the exclusion of each variable
- find the variable with the lowest *F* statistic and eliminate this if the statistic is smaller than some predetermined value
- This leaves a model with one fewer variable. Now fit this model and re-run the process above.
- Stop when a variable is not omitted (because the smallest *F* statistic is no longer smaller than the predetermined value).

5.3.2 Stepwise Regression

Stepwise Regression, sometimes called Modified Forward Regression works in the opposite direction to Backward Elimination. This process can be summarised as

- Start with the null model $\beta_0 + \varepsilon_i$
- fit simple linear regression models for each of the explanatory variables under consideration
- select the explanatory variable whose simple linear regression model has the largest *F* statistic (or *t* statistic)
- now add the explanatory variable with the next highest *F* statistic
- test (via subset deletion) whether either of the two variables can be omitted according to some predetermined *F* value. [Sometimes the process may have a higher value needed for omission of an existing variable than for the newest variable just added]
- continue until no more variables are added or omitted.

One problem with this approach is the estimation of σ^2 in the *F* tests. By starting with the simpler models, the MS_E model estimate of σ^2 is likely to be higher in the early rounds of this process and then fall over time as more variables are added. This distorts early round *F* statistics compared to later ones. A potential way around this is to use the full model MS_E as the estimator of σ^2 in all the *F* tests beginning with the first explanatory variable to be added to the null model.

Another variation of the stepwise process is one that only has addition and not omission of existing variables, that is once an explanatory variable is added it cannot then be omitted later in the process.

5.4 Akaike's Information Criterion

One of the main issues with automatic selection processes is the risk that we keep or include explanatory variables whose parameter values are really zero, that is we fail to reject H_0: $\beta_j = 0$ for some j when we should have rejected it [sometimes this is known as a

'false positive'). Akaike's Information Criterion or AIC can help address this. The AIC uses maximum likelihood estimation of parameters that we covered earlier. Akaike's Information Criteria is defined as

$$AIC = 2(p + 1) - 2logL$$

where:

- p = the number of regression parameters (so $p - 1$ explanatory variables)
- L is the Likelihood function evaluated at the maximum likelihood estimates of each of the parameters

We seek the regression model that minimises AIC.

We have already seen in the section on maximum likelihood estimation above that in linear regression models the maximum likelihood estimators of the beta parameters are the same as the least squares estimators.

However the MLE of σ^2 is not our usual unbiased estimator MS_E. Instead the MLE for the model variance is $\hat{\sigma}^2 = \frac{SS_E}{n}$.

Now it can be shown that in a normal linear regression model that

$$-2logL = n(log2\pi + log\sigma^2 + 1)$$

If we seek the model that minimises AIC (which because of the $-2logL$term in the AIC is equivalent to the model that maximises likelihood) then once again there are several ways we can go about this analogous to backward elimination and forward stepwise regression.

Backwards, we start with the full model.

- Construct the full model and calculate its AIC
- Construct all the possible models that omit one variable and calculate the AIC for each of them
- Compare the AIC of all the models (full and each one with a variable omitted)
- If the full model has the lowest AIC use that model and stop
- If one of the other models has the lowest AIC move onto that model and repeat by once again considering all models with one less variable and their AIC's
- Stop once AIC can no longer be reduced by omitting a variable.

This process can be automated in R programming using the `step()` function. If the full model is constructed using `lm()` and stored in R as `full_model` the backwards route to a reduced model is given by

```
reduced_model <- step(full_model, direction = "backward")
```

Forwards we start with the null model. In R the null model is found by `lm(y~1)` If we store this as `null_model` and we have six possible explanatory variables to consider adding, `x1 x2 x3 x4 x5 x6` then this is done in R programming with

```
forward_model <- step(null_model, scope = x1+x2+x3+x4+x5+x6,
direction = "forward")
```

A third variation is to set `direction = "both"` in the `step()` command which has the effect of adding additional variables from the null model and then later deleting one or more of those variables once others are added.

Backwards and forwards processes using AIC may lead to different recommended models.

6 Problems fitting multiple regression models

We now turn to some of the potential problems fitting multiple linear regression models. We begin with one that is a particular issue when the number of explanatory variables becomes large.

6.1 Singular or near-singular $\boldsymbol{X^TX}$

We have already seen that for there to be a solution to the normal equations and a unique set of least squares estimators for **β** that we need $\boldsymbol{X^TX}$ not to be singular. If it is singular, its determinant is zero and there is no unique solution to the normal equations.

The problem of singularity in $\boldsymbol{X^TX}$ occurs when there is linear dependence between the x_{ij} variables, for example

- if two or more variables are equal
- if one variable is a linear combination of other variables

For example consider a model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

In the extreme example where $x_{1i} = x_{2i}$ = 1 all *i* =1,2,3 then

If $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

and $\boldsymbol{X^TX} = \begin{pmatrix} 3 & 3 & 3 \\ 3 & 3 & 3 \\ 3 & 3 & 3 \end{pmatrix}$

$\det(\boldsymbol{X^TX}) = 0$

Or where $x_{1i} = x_{2i}$ but the individual observations take different values

If $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 3 \end{pmatrix}$

then again $\det(\boldsymbol{X^TX}) = 0$

which means that $\boldsymbol{X^TX}$ is not invertible and we cannot solve the normal equations.

Or where x_{2i} itself has a linear relationship to x_{1i}. So if $x_{2i} = 2x_{1i}$ e.g.

If $\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \end{pmatrix}$ then again $\det(\boldsymbol{X^TX}) = 0$

These cases are generally quite easy to identify and eliminate. In modelling scenarios a more common situation is where one explanatory variable's observation set is very close to that of another (or a linear transformation of other variables). Then the determinant of $\boldsymbol{X^TX}$ will be close to (but not equal to) zero.

for example

If $\boldsymbol{X} = \begin{pmatrix} 1 & 0.95 & 1.04 \\ 1 & -0.98 & -1.01 \\ 1 & 1.03 & 0.96 \end{pmatrix}$ then $\det(\boldsymbol{X}^T\boldsymbol{X}) = 0.1014$

In this case $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 78.23 & 0.13 & -78.04 \\ 0.12 & 0.38 & -0.01 \\ -78.04 & -0.01 & 78.23 \end{pmatrix}$

But we know that $var(\hat{\beta}_j) = \sigma^2 c_{jj}$ where the c_{jj} are taken from the diagonal of the matrix above. Therefore in this example $var(\hat{\beta}_2) = 78.23\,\sigma^2$ which is very large in comparison to σ^2.

Parameters with large variances is one of the problems of *multicollinearity*, where some of the columns of $\boldsymbol{X}$ are or are close to linear combinations of other columns. When the variance is very high this can even lead to a parameter having the wrong sign. That is a parameter that should be positive because when the explanatory variable increases, the response should increase as well, is in fact negative (or vice versa). Multicollinearity can also lead to issues with selecting a subset of variables in a multiple linear regression model.

When the number of explanatory variables is relatively small it may well be possible to spot multicollinearity by scanning the data. However as the number of variables increases this may not be possible and we need to develop statistical techniques for identifying potential problems.

6.2 Variance Inflation Factor

The Variance Inflation Factor or VIF is one way to detect possible multicollinearity.

Consider a multiple linear regression of y_i on $p - 1$ explanatory variables, such that the model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i$$

Now the VIF can be calculated for each of the explanatory variables x_j $j = 1, 2, \ldots p - 1$ as follows.

- perform a multiple linear regression of x_j against the other $p - 2$ explanatory variables (so instead of y as the response, x_j is the response and the remaining x's are explanatory variables with their own set of β parameters different from the original regression of y)
- calculate the coefficient of determination of this regression of x_j and express it as a real number between 0 and 1 not a percentage, called R_j^2.

then

$$VIF_j = \frac{1}{1 - R_j^2}$$

Large R_j^2 (close to 1) indicates a strong linear relationship between x_j and other explanatory variables and will lead to a high VIF.

We usually take $VIF_j > 10$ as an indication that multicollinearity might cause problems with a regression model. Where this is detected, we need to simplify the model by reducing the number of explanatory variables so that linear combinations are removed.

Another indication of multicollinearity can be a model where the overall model shows significance with an *F* test but none of the individual parameters show significance with their *t* tests.

6.3 Residuals and their plots

We have already defined the residuals in a multiple linear regression using the matrix form and stated some of their basic properties.

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

$$E[\boldsymbol{e}] = \boldsymbol{0}$$

$$var(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

where ***H*** is the hat matrix

If we denote the (i,j)th element of the hat matrix h_{ij} so that the diagonal elements are h_{ii} then we can re-write the variance of the residuals as

$$var(e_i) = (1 - h_{ii})\sigma^2$$

$$cov(e_i, e_j) = -h_{ij}\sigma^2$$

Note that in the simple linear regression model we referred to h_{ii} as v_i and either notation can be used in multiple linear regression.

The formula for the variance of each residual in a multiple linear regression model gives us an additional reason for standardising the residuals. In simple linear regression we standardised because the variance of the residuals was different to the σ^2 assumed for the random error terms in the original model specification. Now in multiple linear regression with

$$var(e_i) = (1 - h_{ii})\sigma^2$$

we can see that not only is that still the case but furthermore the variance of each of the residuals might be different depending on the hat matrix. This can make detection of outliers difficult, so again it is helpful to standardise residuals calculating d_i where

$$d_i = \frac{e_i}{\sqrt{S^2(1-h_{ii})}}$$

If the normal distribution assumption for the residuals is followed, then $d_i \sim t_{n-p}$

Furthermore for large number of observations *n* the value of h_{ij} (i ≠ j) tends to be small and therefore asymptotically the standardised residuals d_i are iid N(0,1). It is this property that we rely upon the most for model checking using d_i and so we do well to remember this carries most validity when we have a large sample size of observations data.

Our four most common checks using the standardised residuals are similar to those for simple linear regression models:

- plot d_i against each of the explanatory variables x_j (j = 1, 2, … p – 1) individually to check for a linear relationship in that explanatory variable. We seek a plot that does not have a clear pattern, in particular one that does not exhibit clear curvature.
- plot d_i against the fitted values $\widehat{y_i}$. This is to check the assumption of a constant variance (σ^2). Again we seek a plot without a clear pattern. This time the pattern we are particularly wary of is a "funnel" shape indicating increasing variance with y.
- the QQ plot is used to check the assumption of normally distributed residuals. We seek a plot close to the straight QQ line. A QQ plot that deviates from the line is evidence that a transformation of the response may need to be considered.
- Any of the above plots can be used to detect outliers, those observations with particularly large absolute residuals.

Where observation data is such that we record time, it is also useful to plot the standardised residuals in order against time *t*. This can be used to detect "autocorrelation" and is covered in detail in MTH6139 Time Series.

6.4 Influential observations and leverage

We previously discussed influential and high leverage observations in simple linear regression models. Previously with simple linear regression we calculated the leverage of an observation and labelled this v_i. Now with the matrix approach to linear regression we can relate leverage to the hat matrix ***H***. In fact $v_i = h_{ii}$ the leverage of the ith observation set is equal to the ith element on the diagonal of the hat matrix.

The fitted model is $\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{HY}$ and using the hat matrix in the calculation of fitted values given observations of the response leads to the following equation for the ith fitted value

$$\widehat{y_i} = \widehat{\mu_i} = \sum_{i=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{i \neq j} h_{ij}y_j$$

So h_{ii} indicates the extent to which the observation with $y_i$contributes to the fitted value $\widehat{\mu_i}$. This is its leverage. When we think of leverage as the i[th] diagonal element of the hat matrix rather than a separate calculation *v*, a number of properties of h_{ii} emerge.

- Previously we noted that $var(e_i) = \sigma^2(1 - h_{ii})$. Now $h_{ii} < 1$. But h_{ii} close to 1 will give $var(e_i)$ close to zero, that is a fitted value close to the observed value.
- In general h_{ii} is small when x_i is close to its mean $\bar{x}$ and gets larger the further x_i is from its mean.
- $\frac{1}{n} < h_{ii} < 1$ and $\sum_{i=1}^{n} h_{ii} = p$ (the number of parameters). In the simple linear regression model we had sum of leverage = 2 and now this is the general case for *p* parameters and *p* – 1 variables. This means that average leverage is p/n. Thus we usually consider leverage > $2p/n$ as "high leverage" and > $3p/n$ as "very high leverage". There are a number of possible causes of high leverage including the method of data collection used or a single non-typical observation.

We measure leverage and check for high leverage because we are concerned to check whether a single observation exerts influence over the model results. A measure of influence can be obtained from Cook's Statistic. For multiple linear regression and the matrix approach, generalising the formula for Cook's Statistic given previously for simple linear regression we have

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^{\boldsymbol{T}}(\boldsymbol{X^T X})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{p\, S^2}$$

where

$\widehat{\boldsymbol{\beta}}$ is the vector of least squares parameters

$\widehat{\boldsymbol{\beta}}_{(i)}$ is the estimates of the parameters found when the i[th] observation is omitted.

Once again, an unusually large value for D_i can be taken as evidence of an influential observation.

7 Linear Models

In this module we have been concerned with *linear* models, in particular the simple and multiple linear regression models. But what about these models is linear? Perhaps not what many people might naturally assume. A linear model is one that is linear in the parameters (not one that is linear in the explanatory variables). These are all examples of linear models:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \sqrt{x_{2i}} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 \sin(x_{1i}) + \beta_2 x_{2i} + \varepsilon_i$$

Now sometimes (but not always) a non-linear model can be converted into a linear one through a transformation of the response. We say that the model is *linearised*. For example,

$$y_i = \varepsilon_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$$

is not linear, but can be linearised by taking natural logarithms so that

$$\ln(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ln(\varepsilon_i)$$

Note that if we do this, we need to take care about the assumption we make for the distribution of the residuals. To use the techniques we have developed for assessing linear regression models in this module we will need to assume $\ln(\varepsilon_i) \sim N(0, \sigma^2)$ for some constant variance σ^2 and not that $\varepsilon_i \sim N(0, \sigma^2)$ as has been the case before now.

Furthermore other variations of this example such as

$$y_i = \varepsilon_i \exp(\beta_0 + \beta_1 x_{1i} + \frac{\beta_2}{x_{2i}})$$

can also be linearised by a log transformation of the response.

The non-linear model

$$y_i = \alpha + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i)$$

where α is a constant, can also be linearised, this time by subtracting the constant and then taking logs

$$\ln(y_i - \alpha) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Another model not linear in its parameters is

$$y_i = \frac{1}{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i}$$

This can be linearised by inverting the response as long as we are prepared to accept the condition that $y_i \neq 0$

Then the linearised model is

$$\frac{1}{y_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

A particular set of linear models that can be useful in some contexts are *Polynomial Regression Models.*

For example the *Quadratic Regression Model* is written

$$y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$$

This is a linear model because it is linear in β_1 and β_{11}

In some ways it might be better to think of the quadratic model not so much as a multiple linear regression model but as an extension of the simple linear regression model

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

That is because the quadratic model only has one explanatory variable, but that variable appears twice as x_i and x_i^2.

We can compare the simple linear and quadratic models for a certain set of observations. Having fitted $y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$ we can test

$H_0 : \beta_{11} = 0$ versus $H_1 : \beta_{11} \neq 0$

Under H_0 a simple linear regression model adequately describes the data whereas if we reject H_0 then the quadratic component gives a statistically significantly better fit.

We can extend this to cubic and higher order polynomials in x_i However for most data sets, higher powers of x_i quickly become very large. Therefore it is often sensible to centre the data and model $z_i = x_i - \bar{x}$ instead of x_i.

We can also have polynomial regression with multiple explanatory variables where each explanatory variable appears multiple times in the model with different powers. For example with 2 explanatory variables x_1 and x_2, and allowing quadratic terms in each, we get the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{1i}^2 + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$$

If we were considering subset deletions in these types of models it is usual not to remove the first order terms x_{1i} or x_{2i} whilst leaving any of the higher order terms that rely upon them still in the model.

These types of models are important in certain applications and are sometimes called Response Surface methods (RSM). This gives a method for statistical modelling of variables that are often represented geometrically by a three-dimensional surface. This has proven a useful way of analysing results of some experiments in biology and chemistry.