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

Now the observed values $y_1, y_2, \dots\ y_n$ will be different to the fitted values $\hat{y}_1, \hat{y}_2, \dots\ \hat{y}_n$ 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 $\sum_{i=0}^{n} e_i = 0$

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

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*(β_0, β_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*(β_0, β_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, …, 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, …, 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 $SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^{\,2} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^{\,2}$ 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 *v₁* degrees of freedom and variable *Y* follows a Chi-squared distribution on *v₂* 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 $H_0\colon \beta_1 = 0$ versus the alternative $H_1\colon \beta_1 \neq 0$ 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 not 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

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

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

So the mean of the ith 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.