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^2_{n-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