Contents

1 *What is a Linear Model ?*

The aim of a statistical model is to predict an *outcome* variable based on one or more *predictor* variables. Outcome and predictor variables have many names: some are given table 1.1. A linear model is a particularly simple statistical model that assumes that the relationship between the outcome and the predictor(s) is a linear one, i.e. it can be described using straight lines.

Outcome	Predictor
Y-variable	x-variables
Dependent variable	Independent variables
Response variable	Regressors
Output variable	Input variables
	Explanatory variables
	Carriers
	Covariates

Table 1.1: Names given to predictor and outcome variables

1.1 Linear Model Equation

Suppose that we have an outcome variable Y and p predictor variables from which we wish to predict Y, and that we have measured Y and x_1, x_2, $\ldots x_p$ on n subjects. In this case, the equation of the linear model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{1.1}$$

The part of the equation $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ is called the *Linear Predictor*. Its value is called the *predicted value of* Y and often written as $\hat{Y}$ (pronounced "Y hat"). This part of the model represents the variation in Y that can be predicted, and may be referred to as the *systematic* component of the model. The term ε is called the *error term* (also referred to as the *noise* or *random* component of the model). This represents the variation in Y that cannot be predicted. We assume that ε has a normal distribution, with mean 0 and variance σ^2_ε. The variance of Y is equal to the sum of the variance of $\hat{Y}$ plus σ^2_ε: in other words the total variability in Y is equal to the predictable variability plus the unpredictable variability. Clearly, we would like the predictable variability to be as great as possible and the unpredictable variability to be as little as possible.

1.2 Linear Model Assumptions

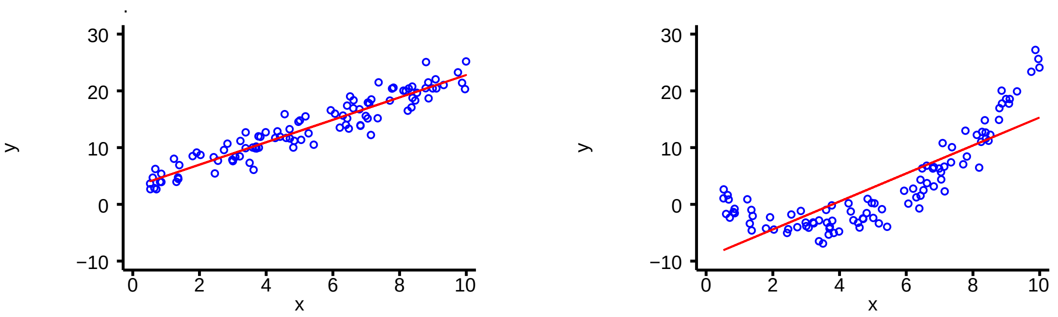
Implicit in the above equation are a number of assumptions about the data. They are

Variables Y_1, Y_2 ... Y_n are independent. Equation 1.1 assumes that only the x-variables are important for predicting Y. If Y depends also on other Y values, (for example in a time-series, when the value of Y may depend on the previous value of Y as well as the x-variables), the importance of the x variables can be over-estimated.

The variance of Y_i is constant. We assume that the error term has the same distribution irrespective of the values of x or Y.

Mean of Y is a linear function of x A linear model fits the best straight line to the data, as shown in Figure 1.1(a). If the true association between x and Y is not a straight line, the model will provide poor predictions. This can be seen in Figure 1.1(b), which illustrates a nonlinear association between x and Y, together with the fitted linear model. Due to the non-linearity, the predicted values are less than the observed values for extreme x-values and greater than the observed values for more central x values.

Distribution of Y_i is normal. This is particularly important in small samples, where outlying observations can have a large effect on the regression line.



(a) Linear Model

(b) Non-linear Model

Figure 1.1: Comparison of Linear and Non-Linear models

Before doing any regression analysis, it is worth plotting Y against each of the x-variables to see if the above assumptions are reasonable. There are better ways to test the assumptions that we will consider later, but this first step will tell you if fitting a linear model is completely unreasonable.

1.3 Linear Model Parameters

The terms β_j in the linear predictor are called the *parameters* or *coefficients* of the regression model[a]. The meaning of the parameters is illustrated in figure 1.2. β_0 is the expected value of Y when all of the x-variables are 0 (the *intercept*). If x_j increases by 1, and all the other x-variables remain unchanged, then the expected value of Y increases by β_j. β_j may be referred to as the *slope* or *gradient* of Y on x_j.

[a] Avoid using the expression "beta-coefficient". Although commonly used to refer to regression coefficients, technically a beta coefficient is derived from data in which the standard deviation of all variables is 1

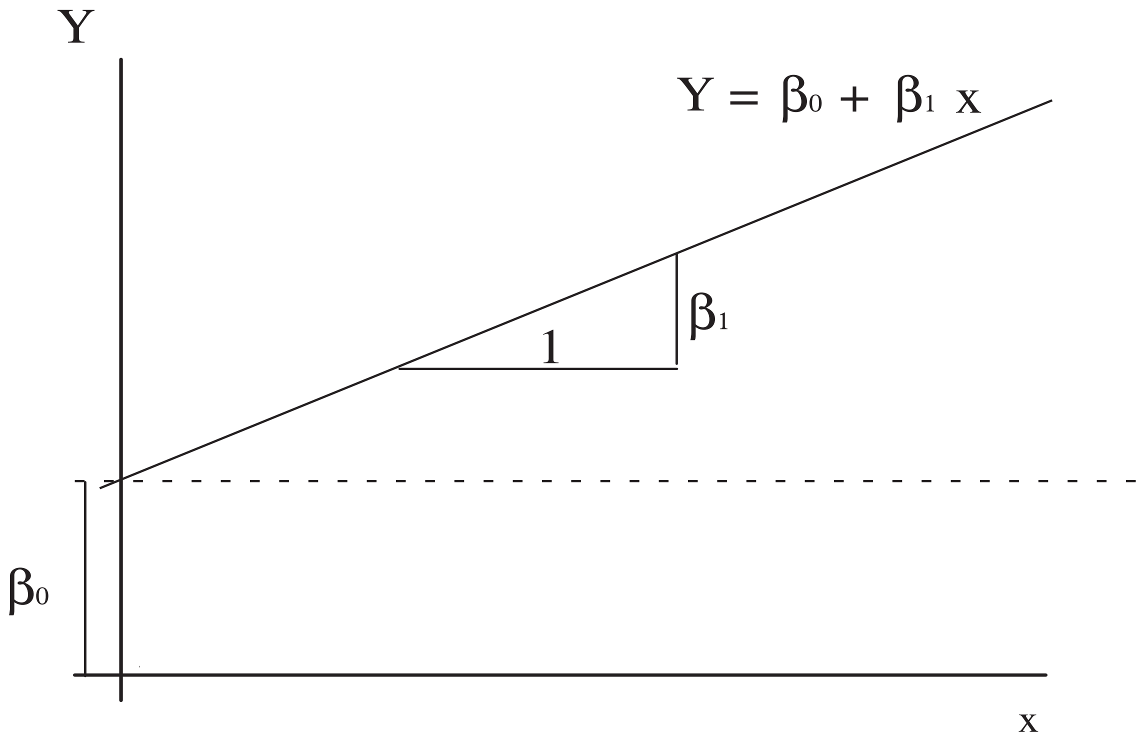


Figure 1.2: Interpretation of Linear Model Parameters

Parameter Estimation

The β and σ parameters describe the relationship between Y and the x-variables in a population. However, we do not know these parameters, we can only estimate them based on a sample from the population. As is customary, we will use greek letters for population parameters and the equivalent roman letter for our estimate of the parameter: b_j is our estimate of the population parameter β_j and s is our estimate of the population parameter σ.

In general, the values given to b_0, b_1, etc. are those which minimise the sum of the squares of the residuals. These estimates are called the "Least Squares" estimates for this reason.

Inference on Parameters

If the assumptions of the linear model are correct, then the parameter estimates b_j will be normally distributed with mean β_j, and standard deviation

$$SD(b_j) = \sqrt{\frac{\sigma^2}{n s_x^2}}$$

where σ^2 is the variance of the error terms ε, s_x^2 is the variance of x_j and n is the number of observations (for sufficiently large n). Note that the standard deviation decreases as n increases (more observations give more precise estimates) and as s_x^2 increases (a wide range of x values gives a more precise estimate than a a narrow range).Since we know the distribution of b_j, we can perform hypothesis tests and construct confidence intervals for β_j.

Unfortunately, we do not know σ, we only have an estimate of it, s_ε, from the data. The slope estimates b_j will therefore follow a t-distribution with $n-p-1$ degrees of freedom, with mean β_j and standard deviation

$$s_{b_j} = \sqrt{\frac{s_\varepsilon^2}{n s_x^2}}$$

This means that we can test hypotheses about β_j using t-statistics. To test the hypothesis that $\beta_j = B$, we form

$$t = \frac{(b_j - B)}{s_{b_j}}$$

This can be compared to a t-distribution on $n-p-1$ degrees of freedom, to provide a p-value for the hypothesis. Most commonly, we wish to test the hypothesis that $\beta_j = 0$, since that would mean that Y is not associated with x_j.

We can also create a $(1-\alpha)$ confidence interval for β_j as

$$b_j \pm t_{\alpha/2,n-p-1} \times s_{b_j}$$

Again, if 0 lies within this confidence interval, we would conclude that x_j and Y are not associated.

Note that if n is sufficiently large, the t-distribution is well approximated by a normal distribution. In this case, a 95% confidence interval can be found by

$$b_j \pm 1.96 \times s_{b_j}$$

However, for small n, the differences between the normal and t-distributions can be considerable, and the above confidence interval will be too narrow. It is therefore better to let stata calculate the confidence intervals for you whenever possible.

The Intercept, β_0

In general, inferences about the intercept, β_0 are of less interest. This is because it is the value taken when all of the x-variables are 0, and usually there is at least one x-variable for which a value of 0 is not sensible (for example height or weight). However, if required, confidence intervals and hypothesis tests can be produced in exactly the same manner as for the slope parameters β_j.

The Outcome Variable, Y

The main purpose of a linear model is to be able to predict a Y-variable from one or more x-variables. Our "best estimate" of Y is given by the linear predictor, but we are also interested in how much individual values of Y will vary around that estimate. There are two sources of variation in Y:

1. The random component of the linear model.
2. Uncertainty about the parameters β_j of the linear model

Clearly, if there is considerable random variation in Y, it will not be possible to predict it with great accuracy. However, it is also clear that the more precisely we have estimated the relationship between x and Y, the more precisely we will be able to predict Y.

Thus the *prediction interval*[b] for Y has the form

$$\hat{Y} \pm t_{\alpha/2,n-p-1} \times \sqrt{s^2_{\hat{Y}} + s^2_{\varepsilon}}$$

The first term under the square root sign represents the uncertainty in the value of $\hat{Y}$, due to the uncertainty about the parameter estimates, the second term represents the random component of the linear model. The second term is assumed to be constant for all values of x, but the first term depends on x: we can estimate Y better in the centre of our data than in more outlying regions. This first term also depends on n: the larger the sample we use, the more precisely we can estimate the parameters of the linear model.

The Fitted Value, $\hat{Y}$

We may be interested in the *mean* value of Y at a given value of x in a population, rather than individual values. In this case, we need not worry about the random component of the linear model, since the mean value of the error terms is 0 by definition. Therefore, a confidence interval for $\hat{Y}$ is given by

$$\hat{Y} \pm t_{\alpha/2,n-p-1} \times \sqrt{s^2_{\hat{Y}}}$$

1.4 Analysis of Variance

We saw above that the total variation in Y is made up of the predictable variation and unpredictable variation. This can be formalised mathematically as

$$\sum (Y - \bar{Y})^2 = \sum \left(\hat{Y} - \bar{Y}\right)^2 + \sum \left(Y - \hat{Y}\right)^2 \tag{1.2}$$

In equation 1.2, the first term is called the *total sum of squares*, since it represents all of the variation in Y about its mean value ($\bar{Y}$). The second term is called *the regression sum of squares*, since it represents the variation of the *predicted values* ($\hat{Y}$) about the mean, i.e. that part of the variation that is predictable. The third term is called the *residual sum of squares*, and represents the variation of the observed Y values about their predicted values. This represents the unpredictable or random variation in Y.

Each sum of squares has an associated *degrees* of freedom (d.f.). The d.f. for the total sum of squares is $n-1$, since the variance of Y is $\sum (Y - \bar{Y})^2/(n-1)$. The d.f. for the regression sum of squares is the number of parameters in the regression model. The residual degrees of freedom is found by subtracting the regression d.f. from the total d.f. This enables us to draw up the following table, called an *Analysis of Variance* or ANOVA table:

The ratio MS_{reg}/MS_{res} is a measure of how much more of the variation in Y is explained by the x-variables than would be expected by chance. If there is no association between Y and the x-variables, this we would expect this ratio to be equal to 1. If the ratio is greater than 1, this suggests that there is an association between Y and the x-variables. MS_{reg}/MS_{res} can be compared to an F distribution to test the hypothesis that there is no association.

If the linear regression model contains only a single variable, then the p-value resulting from the hypothesis test that $\beta = 0$ will be *exactly* the same as that resulting from the hypothesis test that $F = 1$. If there are several predictors, the F-test provides a test of the overall model, whilst the t-tests provide tests of each individual predictor variable.

[b]This is *not* a confidence interval, because only parameters have confidence intervals, and Y is a random variable, not a parameter. In fact, it is a reference range for Y, conditional upon x.

Source	df	Sum of Squares	Mean Square	F
Regression	p	SS_{reg}	$MS_{reg} = \dfrac{SS_{reg}}{p}$	$\dfrac{MS_{reg}}{MS_{res}}$
Residual	n-p-1	SS_{res}	$MS_{res} = \dfrac{SS_{res}}{(n-p-1)}$	
Total	n-1	SS_{tot}	$MS_{tot} = \dfrac{SS_{tot}}{(n-1)}$	

Table 1.2: ANOVA Table

1.5 Goodness of Fit

The fact that there is an association between x and Y does not necessarily imply that x is useful for predicting Y. The statistical significance of an effect depends both on the size of the effect and the size of the sample in which it is being measured. A small effect may be highly significant in a very large sample, but still provide little predictive power.

The predictive power of the linear model depends on how much of total variation in Y can be predicted. We have seen that the predictable variation is SS_{reg} and the total variation is SS_{tot}, so the proportion of the variation that we can predict is

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$

This quantity is called R^2 because it is the square of the coefficient of correlation between Y and $\hat{Y}$: since it is a proportion it can take any value from 0 to 1.

This gives an overall measure of how good a model is, but it is difficult to use to compare models. This is because adding an extra variable to a model will *always* increase R^2, whether or not the variable is related to Y. A better statistic for comparing models is the adjusted R^2, which allows for the fact that even unrelated variables will explain some of the variance of Y. The larger the sample size, the smaller the difference between R^2 and adjusted R^2.

1.6 Linear Models in Stata

The `regress` Command

Linear models are fitted in stata using the `regress` command. The syntax for this command is simply

```
regress yvar xvars
```

The regression can be performed in different groups using `by` or `if` clauses, and there are various other complex options which you need not worry about yet.

Fitted values from a regression model can be obtained by using the `predict` command. Typing

```
predict varname, xb
```

will create a new variable called *varname*, containing the fitted values[c].

The `predict` command can also generate other variables which may be of interest. For example, the standard error of the forecast, $\sqrt{s^2_{\hat{Y}} + s^2_{\varepsilon}}$, can be calculated by

```
predict varname, stdf
```

The standard error of $\hat{Y}$, called the standard error of the prediction in stata can be calculated by

```
predict varname, stdp
```

These enable us to construct confidence intervals and prediction intervals.

Understanding Stata Output

In this section, we are going to use stata to fit a linear model to the data in table 1.3.

x	Y
4	4.26
5	5.68
6	7.24
7	4.82
8	6.95
9	8.81
10	8.04
11	8.33
12	10.84
13	7.58
14	9.96

Table 1.3: Data for Regression

If this data is entered into stata, and we type

```
scatter Y x
```

we get the scatterplot shown in figure 1.3. This scatterplot suggests that Y increases as x increases, and fitting a straight line to the data would be a reasonable thing to do. So we can do that by typing

```
regress Y x
```

The output from the `regress` command comes in two parts: first an ANOVA table and some overall tests of the model, then a table of statistics for the individual coefficients. Here is the first part of the output for the above regression.

[c]The odd name "xb" is derived from matrix algebra, since $b_1x_1 + b_2x_2 + \ldots + b_px_p$ can be written as $\mathbf{xb}$, where $\mathbf{x}$ is the $1 \times k$ matrix $x_1, x_2, \ldots, x_p$ and $\mathbf{b}$ is the $k \times 1$ matrix $b_1, b_2, \ldots, b_p$.

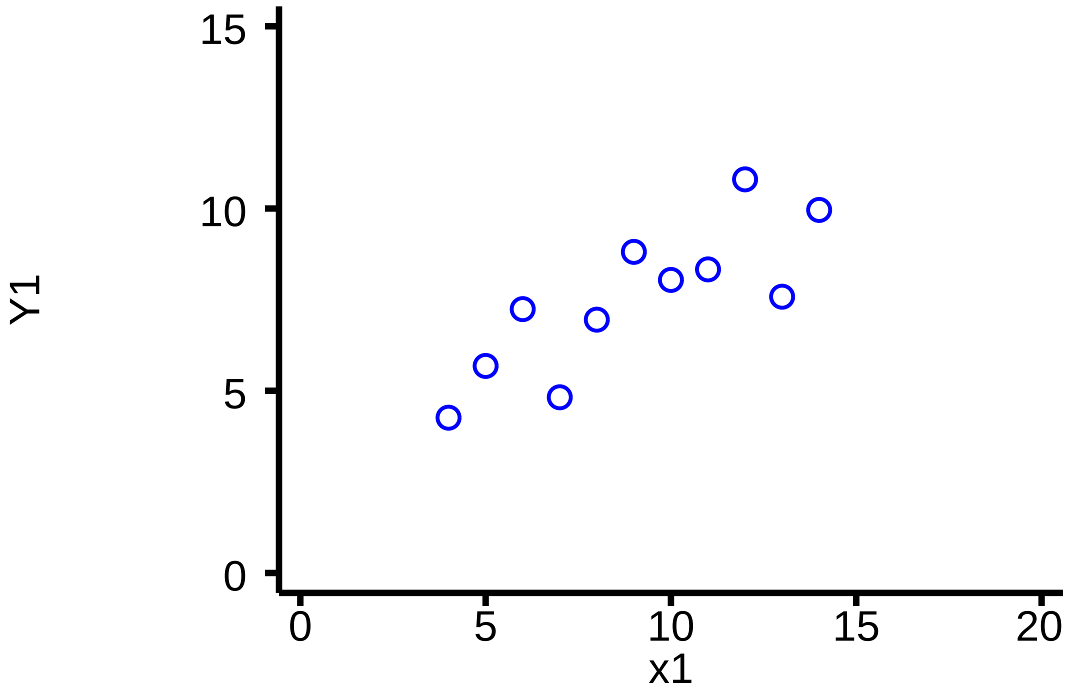


Figure 1.3: Scatterplot of sample linear model data

```
  Source |       SS       df       MS              Number of obs =      11
---------+------------------------------           F(  1,     9) =   17.99
   Model |  27.5100011     1  27.5100011           Prob > F      =  0.0022
Residual |  13.7626904     9  1.52918783           R-squared     =  0.6665
---------+------------------------------           Adj R-squared =  0.6295
   Total |  41.2726916    10  4.12726916           Root MSE      =  1.2366
```

Most of the ANOVA table is on the left, with other statistics on the right. The F value is 17.99, which is considerably greater than 1. The p-value, labelled `Prob > F`, is 0.0022: in other words, if there were no true association between x and Y, only 2 times out of 1000 would random sampling produce such a strong observed association. We can therefore reject the hypothesis of no association. R^2 is 0.6665, so nearly 67% of the variation in Y can be predicted from x, and only 33% is random. Finally, the term `Root MSE` is the square root of the residual mean square MS_{res}, which is an estimator for σ.

Here is the second part of the output which deals with the individual coefficients:

```
------------------------------------------------------------------------------
       Y |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   .5000909   .1179055      4.241   0.002       .2333701    .7668117
   _cons |   3.000091   1.124747      2.667   0.026       .4557369    5.544445
------------------------------------------------------------------------------
```

The intercept parameter, β_0, is labelled `_cons` (for "constant") by stata. Thus our prediction equation is

$$\hat{Y} = 3.00 + 0.500 \times x$$

The table also gives standard errors for the parameters β_1 and β_0, and the corresponding t-statistics to test the hypothesis $\beta = 0$. The p-values are given in the 5^{th} column, showing that in this case both parameters are significantly different from 0[d]. Finally, a 95% confidence interval for each parameter is given.

We may now wish to look at the confidence intervals for the fitted values and the prediction intervals. To see the data along with a 95% prediction interval, we could now use the command[e]

```
twoway lfitci Y1 x1, stdf|| scatter Y1 x1, ylab(0(5)15) xlab(0(5)20)
```

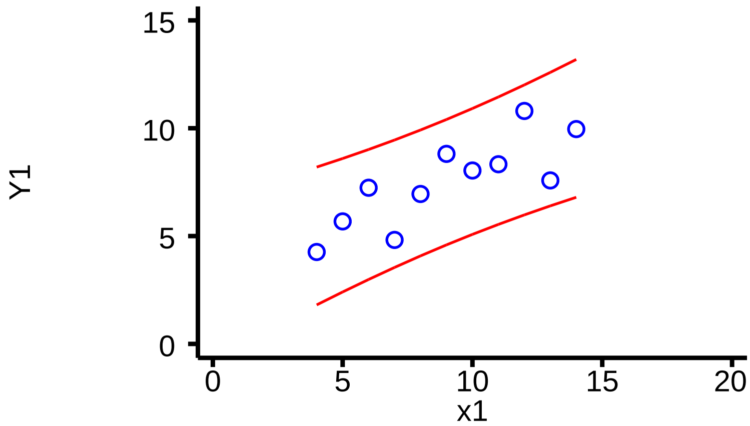
Figure 1.4(a) shows the data along with a 95% prediction interval. We would expect 95% of the observations to lie within this interval: in this case, all 11 observations do. Figure 1.4(b), on the other hand shows the regression line, together with its 95% confidence interval. This figure was obtained by typing

```
twoway lfitci Y1 x1|| scatter Y1 x1, ylab(0(5)15) xlab(0(5)20)
```

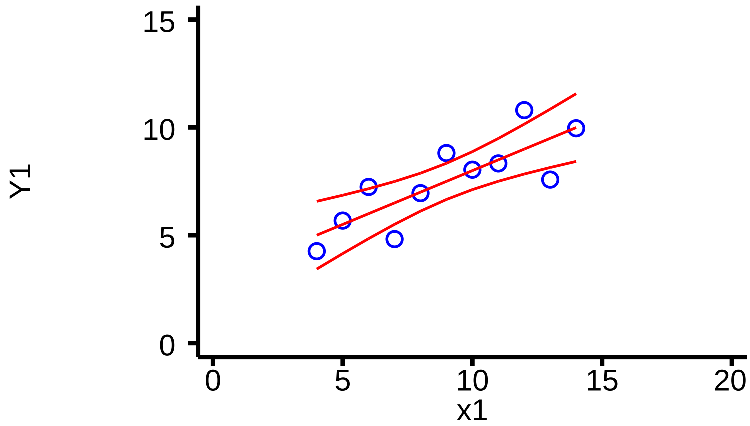
Notice that the confidence interval is much narrower than the prediction interval, since it does not need to incorporate the random element in Y. The confidence interval around the fitted value is analogous to the confidence interval around the mean of sample, whilst the prediction interval is analogous to a reference range for values in the population. Also, both the confidence interval and reference range are narrower near the centre of the range of x-values and wider further away. In particular, the confidence interval and prediction interval are very wide when $x = 0$, since this is well to the left of the data. In fact, the confidence interval when $x = 0$ is (0.46, 5.54), since this is the confidence interval of β_0.

[d]The more observant among you may have noticed that the p-value from the t-test for x is the same as that from the F-test above. This is not a coincidence

[e](The graphics commands will be covered in session 11: for now just take my word for it that this does what it should).



(a) Prediction Interval

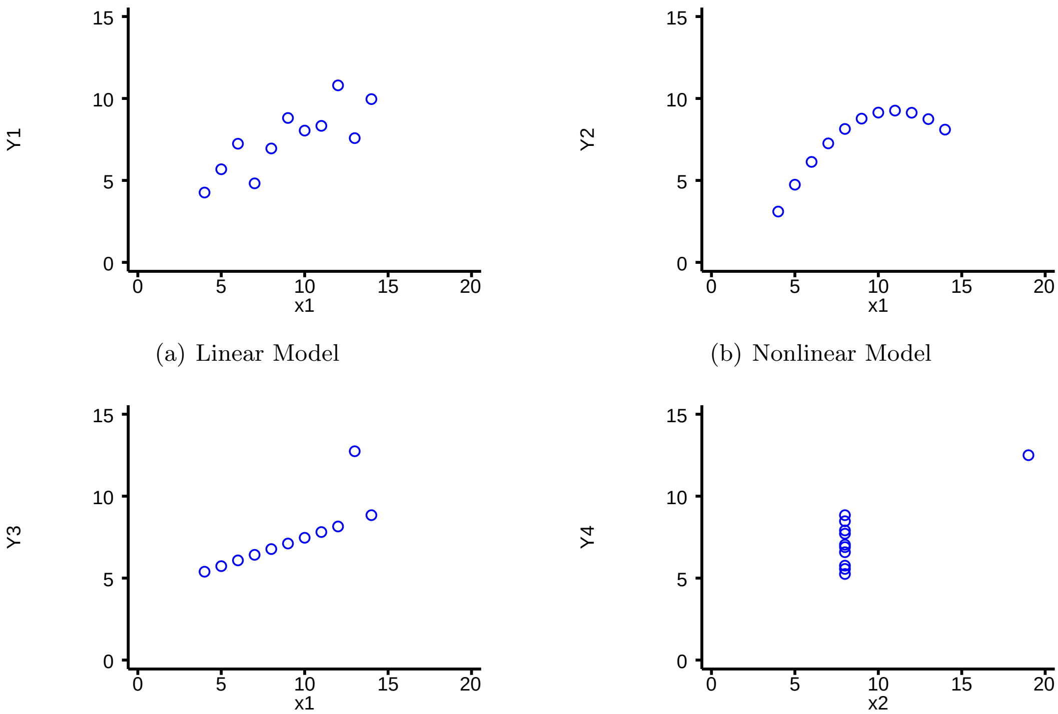


(b) Confidence Interval

Figure 1.4: Prediction Interval and Confidence Interval in Linear Regression

2 Diagnostics

How do we know if the model is adequate ? Part of that question is answered by the R^2 statistic: if R^2 is small, the model is not good at predicting Y. However, there is another aspect to this question: do the data satisfy the assumptions of the linear model outlined in section 1.2. If not, any conclusions drawn may be misleading.



(a) Linear Model

(b) Nonlinear Model

(c) Y Outlier

(d) Y and x Outlier

Figure 2.1: Different Data Configurations Resulting in the Same Linear Model

Consider the scatter-plots in Figure 2.1[a]. In each case, the ANOVA table is the same as that of section 1.6, as are the parameter estimates and standard errors. However, only in figure (a) is a linear model an appropriate fit to the data. Using a linear model to predict values of Y from values of x in the other situations may result in very poor estimates. In the following sections, we will see why a linear model is inappropriate for the other datasets, and how we can recognise similar situations.

[a] These data were devised by F. J. Anscombe, and first explored in “Graphs in Statistical Analysis”, *The American Statistician*, vol 27, pages 17-21.

2.1 Testing Assumptions

When testing that your data satisfies the linear model assumptions, it is important to proceed in the following order:

Confirm Constant Variance If the the variance of Y varies as x varies, a linear model cannot be fitted. There are two possible solutions: either transform Y or use weighted regression, both of which are beyond the scope of this introduction.

Confirm Linearity of Association If the association between Y and x is not linear, this can be solved by either transforming one or more x variables or fitting polynomials x^2, x^3 etc.

Identify Influential Observations If one observation is unusual, it can have an inordinate influence on the regression line, as shown by Figures 2.1(c) and 2.1(d). Such points need to be identified and their influence assessed.

Confirm Normality of ε_i Finally, we should check that the residuals are normally distributed.

2.2 Confirm Constant Variance

The first assumption that must be tested is the constant variance of error terms. This is because if this assumption does not hold, the solution may be to transform Y. If this is necessary, it will change all of the properties of the error terms, so that if they satisfied the assumptions of the linear model before the transformation, they are unlikely to satisfy them afterwards.

To test that the variance of the error terms, $\varepsilon_i = Y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$, is constant, we need to calculate the ε_i. We cannot calculate them directly, because we do not know the values of the β parameters, we only have estimates for them. The obvious estimate for the error term ε_i is $e_i = Y_i - b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi}$. The e_i are referred to as *residuals*.

However, these residuals are not suitable for testing the assumption of constant variance, since even if the ε_i have constant variance the e_i do not. This is because $\hat{Y} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi}$ varies more away from the centre of the data, but Y has constant variance. Therefore, since

$$var(Y_i) = var(\hat{Y}_i) + var(e_i)$$

the variance of the residuals must be *less* at the extremes of the data.

Since we can calculate the expected variance of e_i, we can adjust for it and produce residuals which do have constant variance provided that the ε_i have constant variance. These residuals are called *standardised* residuals (s_i)[b]. I will spare you the mathematical formula: the way to calculate them is to type

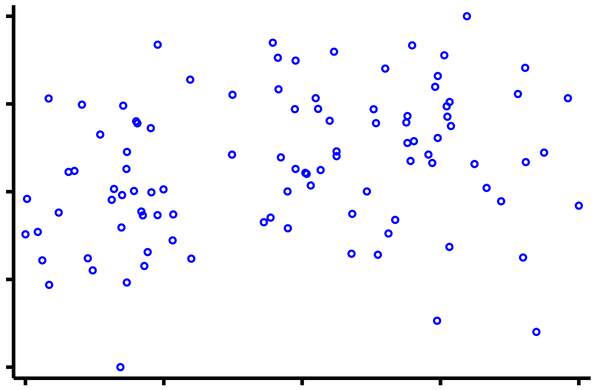
```
predict varname, rstandard
```

after a `regress` command.

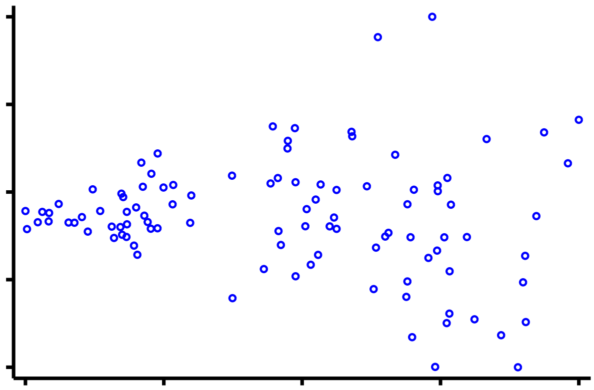
A plot of s_i against $\hat{Y}_i$ should show no pattern. The plot should consist of a rectangular cloud of points, showing that the mean standardised residual is 0 for all values of x, and that the variance of the standardised residuals are the same for all values of x. A typical result of such a plot, is shown in Figure 2.2(a)

There are two patterns that are commonly seen in plots of s_i against $\hat{Y}_i$ which suggest that a linear model is inappropriate. Firstly, the spread of the data may increase (or more rarely

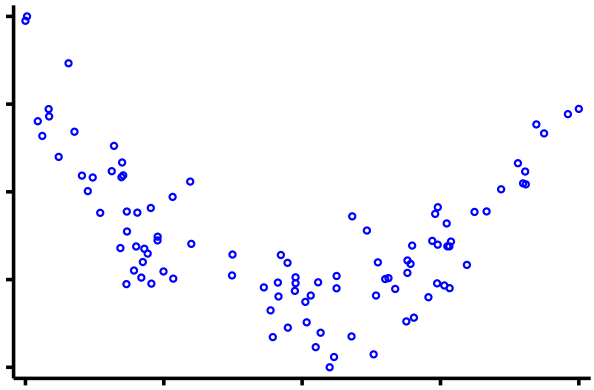
[b] Some statisticians call these residuals *studentised* residuals, but others, including those who wrote stata, use the term studentised residuals for something completely different, which we will come to soon



(a) Normal Distribution



(b) Heteroskedasticity



(c) Non-Linearity

Figure 2.2: Plots of Residuals vs Fitted Values

decrease) as $\hat{Y}_i$ increases. This is called *heteroskedasticity*. It may be possible to remove this by transforming Y. Alternatively, a weighted regression can be used, but this is beyond the scope of this work. This is illustrated in Figure 2.2(b).

Alternatively, the plot may show curvature. This is an indication that the association between Y and at least one of the x-variables is not linear. However, it does not indicate which x-variable is the problem: a different kind of plot (a partial residual plot, described in section 2.3) is required for that. Curvature in the residual vs fitted value plot is shown in Figure 2.2(c).

Stata can produce standardised residuals for you, using the command

```
predict varname, rstandard
```

These can then be plotted against the fitted values produced by

```
predict varname, xb
```

There is a builtin command `rvfplot`, which plots residuals against fitted values, but since it does not use studentised residuals it may be misleading in small datasets. In this case, it is better to to use `predict` to calculate studentised residuals and create the plot yourself. For large datasets, the difference between the residual and the standardised residual is small, and `rvfplot` can safely be used.

If you are unsure whether the plot reveals non-constant variance, there is a command `hettest` to formally test this. Simply typing `hettest` tests whether the variance is a function of $\hat{Y}$, whilst `hettest varname` tests if the variance is a function of the predictor variable `varname`. It is also possible to formally test whether there is curvature in the plot, using the command `ovtest`. However, if there are several x-variables, it is better to consider partial residual plots, outlined in section 2.3

Interpreting the results of the formal tests requires care. If the dataset is very large, it is possible for a very small, unimportant effect to be statistically significant. On the other hand, in a small data set a large effect, which does invalidate the model, may not be statistically significant. Therefore, the plots should always be considered, as well as the results of the formal tests.

2.3 Confirm Linearity of Association

One way to confirm the linearity of the association would be to plot the residuals against each of the predictor variables in turn. However, if there are several predictor variables, it can be more useful to generate the partial residuals

$$p_j = e + b_j x_j = Y - \beta_0 - \sum_{l \neq j} b_l x_l$$

These partial residuals are formed by subtracting from the observed value of Y that part of the predicted value that does not depend on x_j. Thus, a plot of p_j against x_j will reveal the association between Y and x_j after adjusting for the other predictors.

It has been suggested that it is easier to interpret a partial residual plot if $b_j x_{ij}$, is plotted along with it. The term $b_j x_{ij}$ is called the *component*, since it is the component of Y that can be predicted from x_j. This shows the presumed linear association, to make departures from linearity easier to spot. Such plots are known as *component-plus-residual* plots (CPR plots) or *component and component-plus-residual* plots , since the partial residual, p_j is the sum of the component of Y that is due to x ($b_j x_j$) and the residual e.

These plots can be obtained from stata using the command `cprplot` *`varname`*, after having run a regression. They are unnecessary if there is only a single predictor variable (a plot of Y against x provides the same information in this case), but can be very useful if there are a number of correlated predictor variables.

Figure 2.3 shows a CPR plot for the data in figure 2.1(b). It clearly illustrates the curvature in the relationship between Y and x.

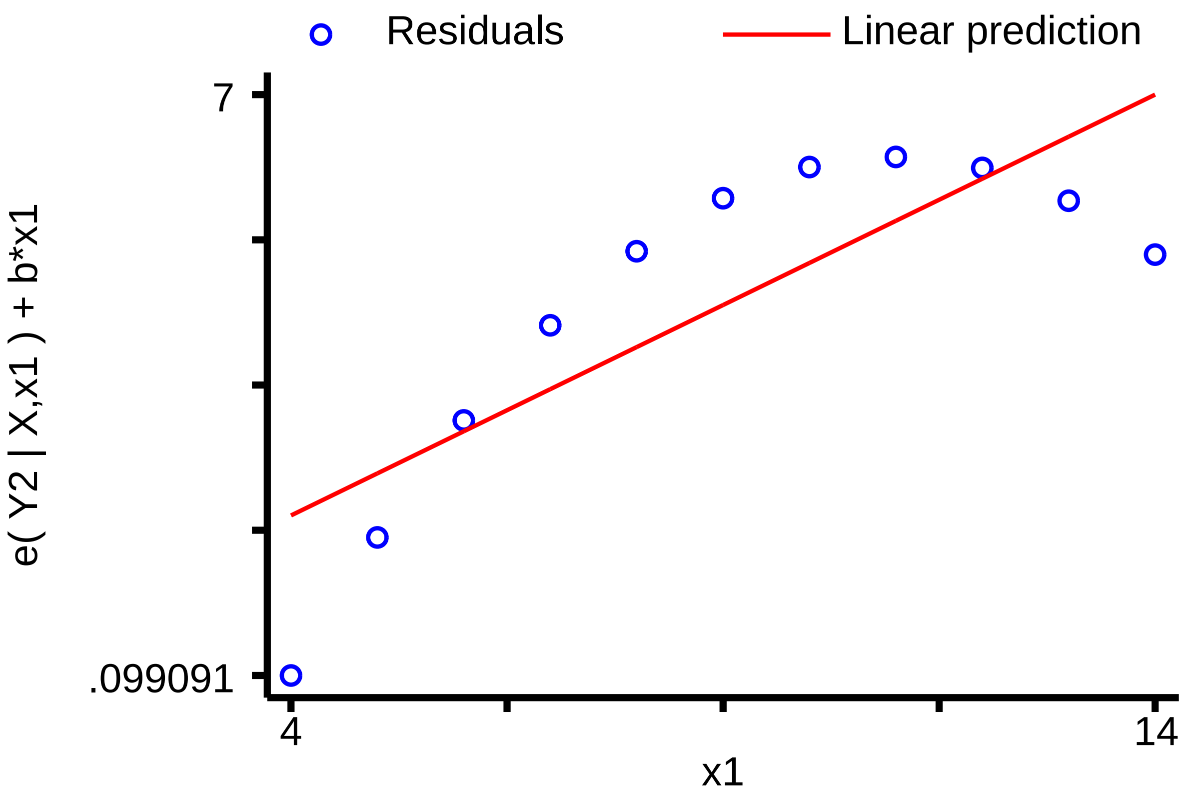


Figure 2.3: Component and component-plus-residual plot for data from Figure 2.1(b)

2.4 Identify Influential Observations

We have seen from figures 2.1(c) and 2.1(d) that a single unusual observation can have a marked effect on the regression equation. Such points are called *influential* points. Identifying influential points can be surprisingly difficult, particularly if there are a number of them and several predictor variables.

The influence a point has on the regression model depends on two factors: how unusual the values of its x-variables are and how unusual the value of its y-variable is. For example, in Figure 2.1(d), the single point with an x value different from the other points completely determines the regression equation. Points with unusual x values are said to have high *leverage*, and are potentially influential points. If, in a addition, the Y value is also unusual, the point will be influential.

2.4.1 Identifying Influential Observations Graphically

One way to detect influential points is to produce a plot of the leverage against the squared residual, since the squared residual measures how unusual the Y value is. This can be produced simply in stata by typing `lvr2plot`. Points near the top of this diagram have high leverage, and are therefore potentially influential. Points to the right of this plot have large residuals. However, influential points, by definition, pull the regression line towards themselves and hence reduce their residuals. For example, in Figure 2.1(d) the point on the right is highly influential because of its x value, but has a residual of 0, precisely because it is so influential. It may be best to re-run the regression excluding any points with high leverage, to see if the conclusions change. Alternatively, there are some statistics that can be used to assess influence in section 2.4.2.

To detect influential points, *studentised* residuals are often used. These are similar to standardised residuals, the regression line from which the residual of a particular point is measured is fitted *excluding that point*. This is because an influential point will attract the regression line towards it, and may therefore have a small standardised residual, but since it is far from the pattern of the rest of the data will have a large studentised residual.

2.4.2 Identifying Influential Observations Formally

There are a number of statistics that can be calculated to determine how influential a point is, depending on what the point may have an influence on. For example, if you are concerned that an observation is having an undue influence on the regression coefficients, you may use the DFBETA statistic. DFBETA_{ij} is a measure of how much β_j is changed if observation i is not included in the regression.

However, in general we are more interested in the predicted values. If there are a number of correlated predictors, an observation may have considerable influence over several coefficients without influencing the predicted value greatly (if the changes "cancel out"). In this case, it may be better to consider DFFITS_i, which gives a measure of the change in the predicted value of observation i if observation i is not used to calculate the regression equation. Another statistic related to DFFITS is Cook's distance. This can be thought of as a measure of the change in *all* the fitted values if an individual observation is omitted.

There is no theoretical basis for deciding when DFBETAS, DFFITS, or Cook's distance are large, since the expected distributions when the data are normally distributed are not known. However, some commonly used cut-offs are given in table 2.1. These should only be taken as suggestions, however, since different authors would choose different cut-offs. An alternative idea is to plot Cook's distance against the predicted value, and look for outlying observations.

Statistic	Cutoff
DFBETAS	$2/\sqrt{n}$, $1.5/\sqrt{n}$
DFFITS	$2\sqrt{(p+1)/n}$
Cook's Distance	4/n

Table 2.1: Commonly Used Cutoffs for Influence Statistics

If the dataset does contain one or more influential observations, what should be done. Firstly, you should check that the unusual observations are not due to an error. If no error can be found, the sensible thing to do is to present the results of fitting the regression both to all points and to the non-influential points, to illustrate the effect of the influential points.

It should also be pointed out that although these statistics are good for detecting individual outliers, if there are several outliers they may not work. It is therefore essential to look at plots of the data to see if there are outliers.

All of the statistics mentioned above can be calculated directly by stata, using the `predict` command after a regression. For details, see the stata manual or type `help regress` into stata.

2.4.3 Y-outliers

Sometimes, there are points which do not have particularly high leverage, but which are influential, due to the observed Y value being very different from the expected value. An example of this is in Figure 2.1(c), where the single outlier pulls the regression line away from the other points, despite not having particularly high leverage.

In this case, there is an alternative to simply deleting the outlying point: robust regression. This involves repeated fitting a regression model, with the weighting of each observation being determined the magnitude of its residual from the previous regression model: the larger the residual, the smaller the weighting (deleting outliers corresponds to giving them a weight of 0, and all the other points a weight of 1). It thus provides a better fit to the bulk of the data, but does not completely ignore outliers.

Figure 2.4 illustrates the use of robust regression on the data in Figure 2.1(c). In this case, the robust regression line passes through 10 of the 11 points in the dataset, and is unaffected by the outlier.

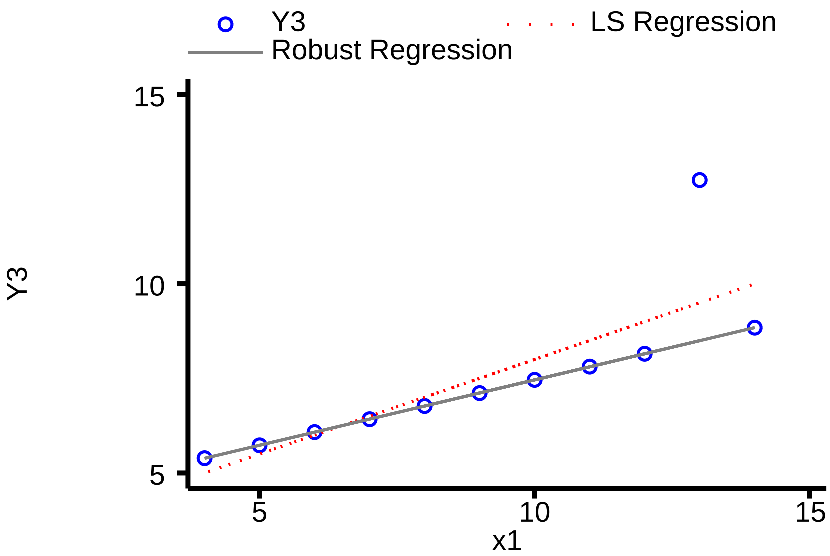


Figure 2.4: Comparison of Least Squares and Robust Regression

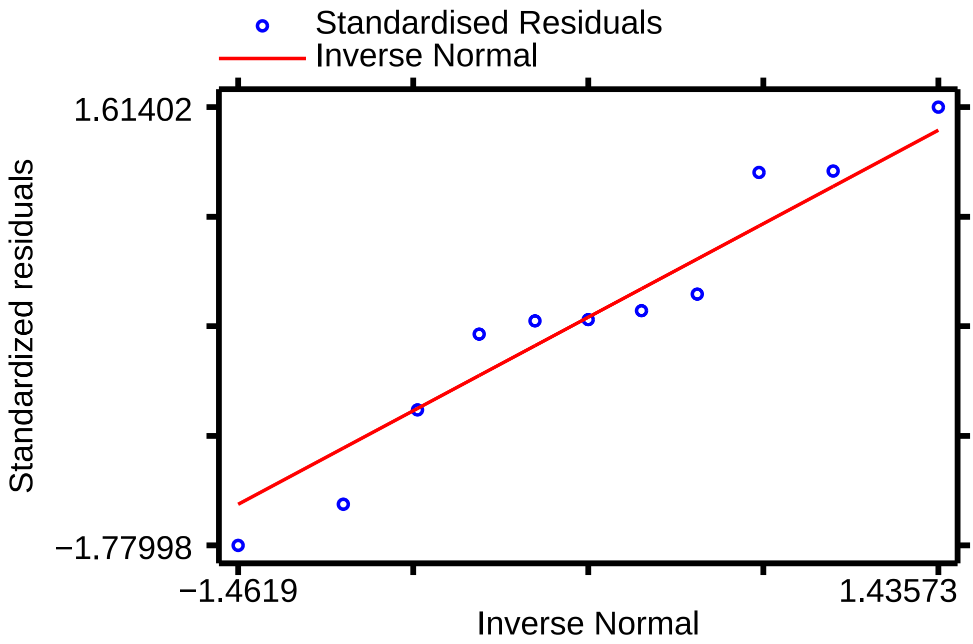
2.5 Confirm Normality of ε_i

Again we use the standardised residuals. A plot of the quantiles of the standardised residuals against the quantiles of a normal distribution should give a straight line: and deviation from a straight line suggests that the residuals are not normally distributed. Such a plot can easily be achieved in stata with the command `qnorm` *`varname`*, where *`varname`* is the name given to the standardised residuals. A formal test of the normality of the residuals is provided by `swilk` *`varname`*.

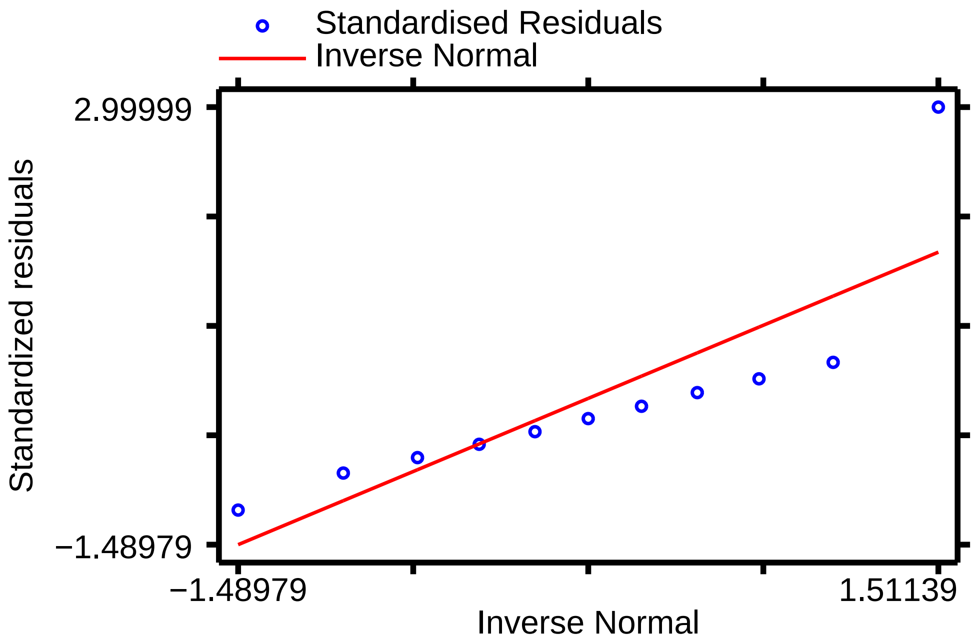
Figure 2.5 below shows normal plots for the residuals in Figures 2.1(a) and 2.1(c). The residuals from the linear model are reasonably normally distributed, but the outlier in the second dataset is clearly visible.

2.6 Dangers of Extrapolation

We have seen that our predictions become less reliable as we move away from the centre of our data. In addition, when we test the assumptions of the linear model, we can only test them within the range of x-values that we have observed. For example, Y may be a linear function of x within the range of x-values measured, but not for more extreme values. For these reasons it is a mistake to predict Y from values of x which lie outside the range of observed x-values.



(a) Linear model



(b) Y and x Outlier

Figure 2.5: Normal plots for the residuals from data in Figures 2.1(a) and 2.1(c).

3 Linear Models Practical

3.1 Datasets

All but one of the datasets that you will use in this practical can be accessed via http from within stata. However, the directory in which they are residing has a very long name, so you can save yourself some typing if you create a global macro for this directory. You can do this by entering

```
global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt
global datadir $basedir/stats/5_LinearModels1/data
```

(That could be entered as a single line, but fitting it on the page would have been tricky). If you wish to run the practical on a computer without internet access, you would need to:

1. Obtain copies of the necessary datasets
2. Place them in a directory on your computer
3. Define the global macro `$datadir` to point to this directory.

The only dataset not stored in `$datadir` is the `auto` dataset, which is distributed with stata. This can be loaded with the command

```
sysuse auto, clear
```

3.2 Fitting and Interpreting a Linear Model

3.2.1 The Anscombe Data

Load the data illustrated in figure 2.1 by entering the command

```
use "$datadir/anscombe.dta"
```

The scatter-plots in figure 2.1 can then be reproduced by entering

```
scatter Y1 x1, xlab(0 (5) 20) ylab(0 (5) 15)
scatter Y2 x1, xlab(0 (5) 20) ylab(0 (5) 15)
scatter Y3 x1, xlab(0 (5) 20) ylab(0 (5) 15)
scatter Y4 x2, xlab(0 (5) 20) ylab(0 (5) 15)
```

The `xlab` and `ylab` options are to ensure that the same axes are used for all graphs. The x-axes go from 0 to 20 with tick-marks every 5, whilst the Y-axes go from 0 to 15 with tick-marks every 5. Note that `Y1`, `Y2` and `Y3` are all plotted against `x1`, but `Y4` is plotted against `x2`.

You should satisfy yourself that these datasets really do produce the same linear models by entering

```
regress Y1 x1
regress Y2 x1
regress Y3 x1
regress Y4 x2
```

3.2.2 The Automobile Data

Load the `auto` dataset which is distributed with stata (the command `sysuse auto, clear` will achieve this). Fit a linear model to predict fuel consumption (`mpg`, miles per gallon) from the weight of the car (`weight`, measured in lbs), using the command `regress mpg weight`

2.1 Is fuel consumption associated with weight ?

2.2 What proportion of the variance in `mpg` can be explained by variations in weight ?

2.3 What change in `mpg` would be expected for a one pound increase in weight ?

2.4 What fuel consumption would you expect, based on this data, for a car weighing 3000 lbs ? (Hint: the command `lincom _cons + 3000*weight` will give you the answer, along with a 95% confidence interval around the expected value. I'll explain how it works next week)

2.5 Would it be reasonable to use this regression equation to calculate the expected fuel consumption of a car weighing 1000 lbs ?

3.3 Diagnostics

3.3.1 Constancy of Variance

Use the data `$datadir/constvar`, which is simulated data generated for this practical.

3.1 Perform a regression of `y` on `x`, using the command `regress y x`. Is there a statistically significant association between `y` and `x` ?

Create standardised residuals using the command `predict rstand, rstand`. Create predicted values using the command `predict yhat`. Now produce a graph of standardised residuals against predicted values with the command `scatter rstand yhat`.

3.2 Would you conclude that the variance is constant for all values of `yhat`, or is there any evidence of a pattern in the residuals ?

3.3 Confirm (or disprove) your answer to the previous question by using the command `hettest`.

3.4 Produce a residual vs fitted value plot with the command `rvfplot`. Would this plot give the same conclusion that you reached in the previous question ?

Use the command `gen ly = ln(y)` to generate a new variable equal to the log of `y`. Perform a regression of `ly` on `x` with the command `regress ly x`. Generate new standardised residuals and predicted values with the commands `predict rstand2, rstand` and `predict yhat2`. Produce a plot of the standardised residuals against the fitted values with `scatter rstand2 yhat2`.

3.5 Is the variance of the residuals constant following this transformation ?

3.6 Confirm your answer to the previous question with the command `hettest`.

3.3.2 Confirming Linearity

Use the data `$datadir/wood73`. This is simulated data to illustrate the use of the CPR plot.

3.7 Plot graphs of `Y` against `x1` and `Y` against `x2` with the commands `scatter Y x1` and `scatter Y x2`. Do these graphs suggest a nonlinear association between `Y` and either `x1` or `x2` ?

3.8 Perform a regression of `Y` on `x1` and `x2` with the command `regress Y x1 x2`.

3.9 Produce CPR plots for `x1` and `x2` with the commands `cprplot x1` and `cprplot x2`. Do either of these plots suggest a non-linear relationship between `Y` and either `x1` or `x2` ?

3.10 Generate a new variable, `x3`, equal to the square of `x1`, with the command `gen x3 = x1^2`. Include `x3` in the regression equation with the command `regress Y x1 x2 x3`. Is `x3` a statistically significant predictor of `Y`l ?

3.11 Produce a CPR plot for `x1`, `x2` and `x3`. Is there still evidence of non-linearity ?

3.12 Use the command `predict Yhat, xb` to generate predicted values for `Y`. Plot `Y` against `Yhat` with the command `scatter Y Yhat`. How good are `x1, x2` and `x3` at predicting `Y` ? Is this what you expected from the value of R^2 from the regression ?

3.3.3 Outlier Detection

Use the data `$datadir/lifeline`. This data was collected to test the hypothesis that the age to which a person will live is governed by the length of the crease across the palm known as the "lifeline" in palmistry. The age at which each subject died is given by `age`, and the length of their lifeline (normalised for body size) is given by `lifeline`.

3.13 Perform a regression of `age` on `lifeline`, using the command `regress age lifeline`. Is there a significant association between age at death and the length of the lifeline in this dataset ?

3.14 Produce a plot of `age` on `lifeline`, using the command `scatter age lifeline`. Are there any points that lie away from the bulk of the data ? If there are, are they outliers in `age`, `lifeline` or both ?

3.15 Are there any points in the above graph that you would expect to have undue influence on the regression equation ?

3.16 Calculate Cook's distance for each observation with the command `predict cooksd, cooksd`. Calculate the predicted age at death for each observation with the command `predict predage`. Plot Cook's distance against predicted age at death with `scatter cooksd predage`. Do any observations have an unusually large Cook's distance ?

3.17 Use `summarize cooksd, det` to identify the value of the largest Cook's distance. Rerun the regression excluding the point with the largest Cook's distance using the command `regress age lifeline if cooksd < x`, for some value of `x` of your choice. How does removing this point affect the regression ?

3.18 Repeat the above analysis removing the two most influential points. Does this change your conclusions about the association between age at death and length of lifeline in this dataset ?

3.19 What is your conclusion about the association between age at death and length of lifeline in this dataset ?

3.3.4 Confirming Normality

We will continue to use the data in `$datadir/lifeline`. Redo the regression including all observations with the command `regress age lifeline`, then use the command `predict rstand, rstand` to produce standardised residuals.

3.20 Draw a normal plot of the standardised residuals with the command `qnorm rstand`. Do the plotted points lie on a straight line ? Are there any observations that do not appear to fit with the rest of the the data ?

3.21 Confirm your answer to the previous question by formally testing for normality of the residuals with the command `swilk rstand`. Do the residuals follow a normal distribution ?

3.4 Complete Example

This example uses `hsng.dta`, a dataset consisting of data on housing in each state in the USA taken from the 1980 census. The variables we are particularly interested in are `rent`, the median monthly rent in dollars; `faminc`, the median annual family income in dollars; `hsng`, the number of housing units; `hsngval`, the median value of a housing unit; and `hsnggrow`, the percentage growth in housing. We are going to see if we can predict the median rent in a state from the data we have on income and housing provision.

Enter the data into stata using the command `use "$datadir/hsng.dta", clear`.

3.4.1 Initial Regression

Use the command `regress rent hsngval hsnggrow hsng faminc` to fit a linear model which predicts `rent` from `faminc`, `hsng`, `hsngval`, and `hsnggrow`.

4.1 How many observations are used in fitting this regression model ?

4.2 How many of the predictor variables are statistically significant ?

4.3 What is the coefficient for `hsnggrow` and its 95% confidence interval ?

4.4 How would you interpret this coefficient and confidence interval ?

4.5 What is the value of R^2 for this regression ? What does this mean ?

3.4.2 Diagnostics: Constancy of Variance

Create standardised residuals using the command `predict rstand, rstand`. Create predicted values using the command `predict pred_val`. Now produce a graph of standardised residuals against predicted values with the command `scatter rstand pred_val`.

4.6 Would you conclude that the variance of the residuals is the same for all predicted values of `rent` ?

4.7 Compare the plot you have just produced to the plot produced by `rvfplot`. Would you have come to the same conclusion about the constancy of variance using this plot ?

3.4.3 Diagnostics: Linearity

Produce a CPR plot for each of the variables `faminc`, `hsng`, `hsngval`, and `hsnggrow`, using the command `cprplot` *`varname`*.

4.8 Is there any evidence of non-linearity in the association between the four predictor variables and the outcome variable ?

3.4.4 Diagnostics: Influence

Calculate Cook's distance for each observation, using the command `predict cooksd, cooksd`. Produce a graph of Cook's distance against the predicted values with `scatter cooksd pred_val`.

4.9 Are there any observations with an unusually large Cook's distance ?

4.10 If so, which state or states ? (You can use the command `list state if cooksd >` *`x`* to find out, by putting in a suitable value for *`x`*.)

Rerun the regression analysis excluding any states with a Cook's distance of greater than 0.5. Use the command

```
regress rent hsngval hsnggrow hsng faminc if cooksd < 0.5
```

4.11 Compare the coefficients and confidence intervals for each of the 4 predictors. Have any of them changed substantially ? Are any of them no longer significantly associated with the outcome ?

Now generate new predicted values with the command `predict pred2`. Compare the new predicted values with the old ones using `scatter pred_val pred2`.

4.12 Is it important to exclude the influential observation(s) from the regression analysis ?

3.4.5 Diagnostics: Normality

4.13 Produce a normal plot of the standardised residuals with `qnorm rstand`. Do the plotted points lie reasonably close to the expected straight line ?

4.14 Use `swilk rstand` to test whether the residuals are normally distributed. Does the result of the test confirm your answer to the previous question ?