

2 Prediction and Analysis of Variance

Reading:

Chapters 1 and 2 of Kennedy *“A Guide to Econometrics”* Achen, Christopher H.

Interpreting and Using Regression (London: Sage, 1982).

Chapter 4 of Andy Field, *“Discovering statistics using SPSS for Windows : advanced techniques for the beginner”*.

Aim:

The aim of this section is to complete our introduction to multiple regression.

Objectives:

By the end of this section you should be able to: understand and apply ANOVA; understand how to use regression for prediction; understand the assumptions underlying regression and the properties of estimates if these assumptions are met.

Plan:

| | | |
|-----|-----------------------------|----|
| 2.1 | Introduction..... | 12 |
| 2.2 | Prediction and Error..... | 13 |
| 2.3 | Errors..... | 16 |
| 2.4 | ANOVA in regression..... | 17 |
| 2.5 | The F-Test..... | 20 |
| 2.6 | Regression assumptions..... | 21 |

2.1 Introduction

One of the most powerful uses of regression analysis is as a means of predicting the values of the dependent variable, sometimes for a future time period, but more generally, for any given set of values of the explanatory variables. This makes it a very powerful tool for policy analysis since a variety of scenarios can be simulated and compared relatively easily and quickly.

Such predictions and simulations, however, are inevitably subject to error, and so analysis of prediction errors is an obvious way of establishing just how well our model fits the data. A straightforward way of doing this is to use the model to predict values for each of the observations in the sample (this is called “in-sample” prediction performance). If we take away the predicted values from the actual values of the dependent variable in our sample, we have what is called “the error term” or “disturbance term”. Hopefully these errors won’t be large.

One way of summarizing the typical amount of error in prediction is to square each of these error values and sum them to give the “Residual Sum of Squares” (if we don’t square them, the negative errors cancel out the positive errors so they would simply sum to zero). In fact, it is this entity – the sum of squared residuals – that is minimised by the algorithm that drives regression (hence the term, “Ordinary Least Squares regression”). This is how a regression draws its line of best fit. Unfortunately, the Residual Sum of Squares on its own is difficult to interpret since it will be contingent on the units of measurement of the different variables in the

regression. Ideally, we would like to have a measure of the goodness of fit that is independent of the scaling of the variables in the regression.

We can arrive at just such a measure by comparing the residual sum of squares with the sum of squared mean deviations of the dependent variable from its mean (termed the “Total Sum of Squares”). The most common way of doing this is to calculate the ratio of the Regression (or “Explained”) Sum of Squares to the Total Sum of Squares, where the Regression Sum of Squares is calculated as the Total Sum of Squares less the Residual Sum of Squares. This ratio is a measure that we have already encountered – it is the Coefficient of Multiple Determination, known more affectionately as simply the “R²”. It is described as the proportion of the variation in the dependent variable (measured by the Total Sum of Squares) to be “explained” by the regression.

It is these topics that form the subject of the present chapter. It is quite a technical area but its worth spending a bit of time mastering the concepts since they are fundamental to later sections.

2.2 Prediction and Error

Given that the regression procedure provides estimates the values of coefficients, we can use these estimates to predict the value of y for given values of x . For example, if the regression output from SPSS is as described in the following table,

| Model | | Unstandardized Coefficients | |
|-------|------------|-----------------------------|------------|
| | | B | Std. Error |
| 1 | (Constant) | -4.200 | 23.951 |
| | X1 | 1.450 | 1.789 |
| | X2 | 2.633 | 3.117 |

a. Dependent Variable: Y

then we can write the equation for the surface of best fit between y , x_1 and x_2

$$y = -4.2 + 1.45 x_1 + 2.63 x_2$$

We can then use this equation to predict the value of y for particular values of x_k . For example, suppose x_1 in the above equation is years of post-school experience, x_2 is years of experience and y is income. Suppose further that we want to know the predicted income of someone with 3 years of post-school education & 1 year experience. We can calculate this by simply plugging in the values for x_1 and x_2 :

$$\begin{aligned} y &= -4.2 + 1.45 \times (3) + 2.63 \times (1) \\ &= \text{£}2,780 \end{aligned}$$

How does this compare with the predicted income of someone with 1 year of post-school education and 3 years work experience? We can answer this by again entering the appropriate values for x_1 and x_2 :

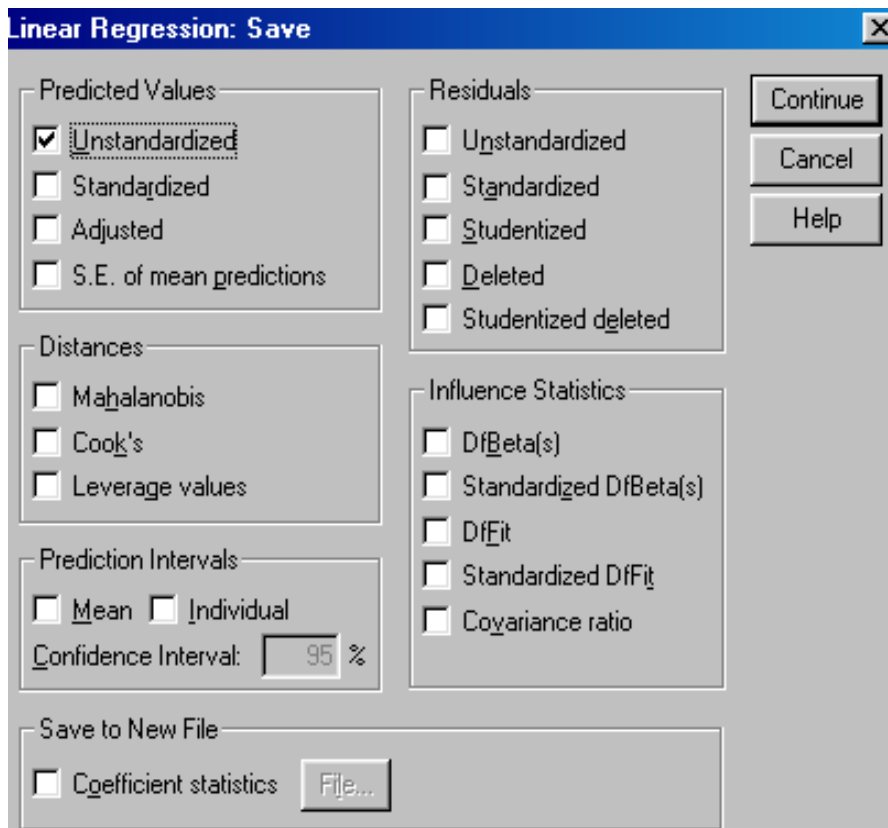
$$y = -4.2 + 1.45 \times (1) + 2.63 \times (3) = \text{£}5,140$$

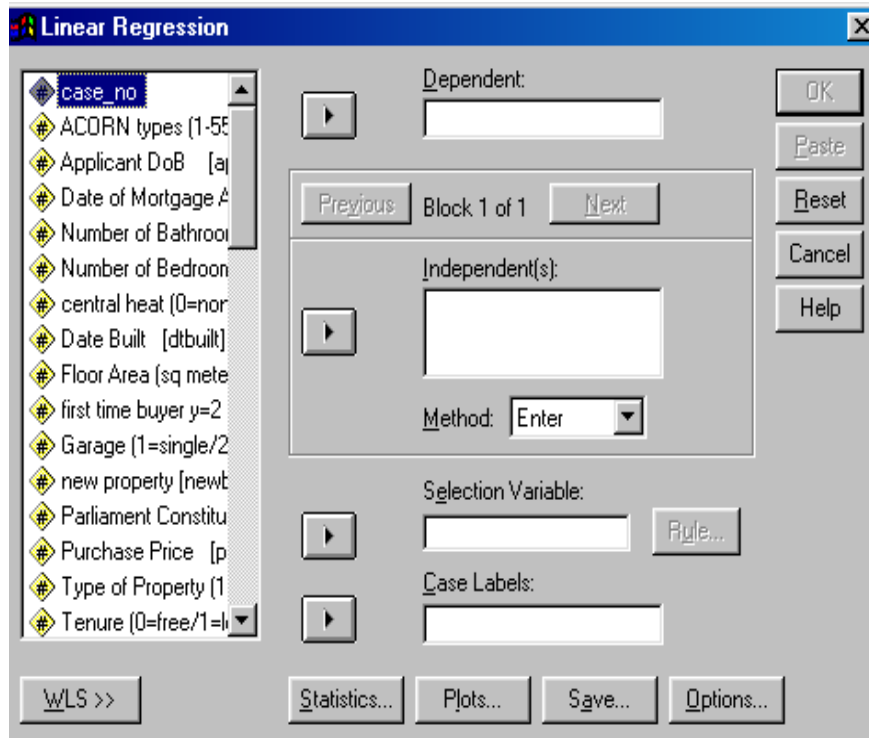
We can also use the equation estimated by regression analysis to predicting the values of y for each value of x_k in the data set. We can either do this by writing a few lines of syntax in SPSS:

```
COMPUTE Y_pred = -4.2 + 1.45*X1 + 2.63*X2.  
EXECUTE.
```

which will create a new variable called “Y_pred” with values calculated for every row in the dataset.

Alternatively, you can ask SPSS to calculate the predicted values for you by clicking on Save in the Linear Regression window, and then selecting the first option on the left hand side: Predicted Values: Unstandardized.





A quicker way of doing this is to add a line to your regression syntax. The normal syntax for running a regression is as follows:

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Y
/METHOD=ENTER X1 X2.
```

The second line “/MISSING LISTWISE” just tells SPSS to ignore observations with missing values for any of the variables in your regression. “/STATISTICS COEFF OUTS R ANOVA” tells SPSS which statistics to list in the output file: “COEFF” = coefficients table, “R” = R-square, and “ANOVA” = analysis of variance. The “CRITERIA” syntax is only applicable if you are using a backwards or forwards elimination procedure to decide on the variables to include (PIN is the significance level at which SPSS includes a variable and POUT is the significance level at which SPSS drops a variable). “NOORIGIN” tells SPSS to include a constant term (i.e. not to run the regression line through the origin). “DEPENDENT” tells SPSS which is the dependent variable and “METHOD” tells SPSS how to enter the variables. The options for METHOD are summarised in the following table, but its worth noting before you consider these that all of the above are the default settings and so if you are happy with these defaults, you can run the same regression by simply writing:

```
REGRESSION /DEPENDENT Y /METHOD=ENTER X1 X2.
```

a) *ENTER Method.*

This is the standard (and simplest) method for estimating a regression equation. All variables for the block are added as a group for the equation. If ENTER is used in a subsequent block, the variables for that block are added as a group for the final model for the preceding block.

b) *REMOVE Method.*

REMOVE is a method that takes variables out a regression analysis. It is used in a block after the first. The variables for the remove block, as a group, are taken out of the final model for the preceding block.

c) *STEPWISE Method.*

This method adds and removes individual variables according to the criteria chosen until a model is reached in which no more variables are eligible for entry or removal. Two different sets of criteria can be used, either,

1) Probability of F. This is the default method. A variable is entered if the significance level of its F-to-enter is less than the entry value (adjustable), and is removed if the significance level is greater than the removal value (adjustable). The entry value must be less than the removal value.

Or,

2) F-Value. A variable is entered if its F value is greater than the entry value, and is removed if the F value is less than the removal value. The entry must be greater than the removal value.

d) *BACKWARD Method.*

This method removes individual variables according to the criteria set for removal until a model is reached in which no more variables are eligible for removal. (If no variables are in the equation from a previous block, they are entered as a group and then removed individually).

e) *FORWARD Method.*

This method adds individual variables, according to the criteria set for entry (see note on c) above), until a model is reached in which no more variables are eligible for entry.

Now if you add “/SAVE PRED” to this regression syntax (before the full stop at the end) as follows:

```
REGRESSION /DEPENDENT Y /METHOD=ENTER X1 X2 /SAVE PRED.
```

then SPSS will automatically create a predicted values variable.

2.3 Errors

Unless your predicted regression line happens to run through all your data points exactly (extremely unlikely), there will be some difference between the predicted value of y from your regression and the actual value you have observed in the data. This difference is called the “residual” or “error term”, e and is calculated by taking away the predicted values of y from the actual values:

$$e_i = \text{prediction error}$$

$$= y_i - y_i^{\text{hat}}$$

Again you could write a couple of lines of syntax to create this variable:

```
COMPUTE e = Y - Y_pred.
EXECUTE.
```

(this syntax assumes that you have already created the Y_pred variable), or you could select it as an option in the “Linear Regression: Save” dialogue box (select Residuals: Unstandardised), or you could include it in your syntax by adding “RESID” to your /SAVE line:

```
REGRESSION /DEPENDENT Y /METHOD=ENTER X1 X2
/SAVE PRED RESID.
```

-
1. Load up Nationwideextract data. Run a regression of purchase price on floor area. Calculate a predicted values variable first using the COMPUTE syntax, then by selecting it as an option in the Linear Regression Window, and then by using the “/SAVE” syntax (give a different name to each version of the variable). Check that they all produce equivalent results by running Descriptive Stats on each of the variables.
 2. Re-run the above regression but this time include number of bedrooms. Calculate the predicted values using the same three methods and compare.
-

2.4 ANOVA in regression

The variance of y is calculated as the sum of squared deviations from the mean divided by the degrees of freedom. Analysis of variance (ANOVA) is about examining the proportion of this variance that is explained by the regression, and the proportion of the variance that cannot be explained by the regression (reflected in the random error term). This amounts to an analysis of the numerator in the variance equation – the sum of squared deviations of y from the mean – since the denominator is constant for all analysis on a particular sample. The sum of squared deviations from the mean is called the “*total sum of squares*” (TSS) and like the variance it measures how good the mean is as a model of the observed data. We can compare how well a more sophisticated model of the data – the line of best fit – compares with just using the mean (where we can think of the mean as “our best guess”).

When a line of best fit is calculated, we get errors (unless the line fits perfectly) and if we square these errors before adding them up we get the “*residual sum of squares*” (RSS). RSS represents the degree of inaccuracy when the line of best fit is used to model the data.

The improvement in prediction from using the line of best fit can be measured by the difference between the TSS and the RSS. This difference is called the “*regression (or*

explained) sum of squares" (REGSS) and it shows us the reduction in inaccuracy of using the line of best fit rather than the mean.

If the explained sum of squares is large then the regression line of best fit is very different from using the mean to predict the dependent variable. That is, the regression has made a big improvement to how well the dependent variable can be predicted. On the other hand, if the explained sum of squares is small then the regression model is not much better than using the mean.

This can be explained as follows: For a particular observation Y_i , the distance from Y_i to the mean of the Y 's, Y^{bar} , can be divided into two parts,

$$Y_i - Y^{\text{bar}} = (Y_i - Y_i^{\text{hat}}) + (Y_i^{\text{hat}} - Y^{\text{bar}})$$

where Y_i^{hat} is the predicted value for Y at that point (ie given a particular X , Y_i^{hat} is the corresponding Y value determined by the OLS line). The distance from Y_i (the observed value) to Y_i^{hat} is called the **residual from the regression**. It is zero if the regression line passes through the point.

The second component ($Y_i^{\text{hat}} - Y^{\text{bar}}$) is the vertical distance from the regression point to the mean of the Y 's. This distance is **explained** by the regression in that it represents the improvement in the estimate of the dependent variable achieved by the regression.

Now suppose we wish to gain some picture of how the **residuals** and **explained** deviations fair for the sample as a whole (ie for all observations). We could just sum $(Y_i - Y^{\text{bar}})$ and $(Y_i^{\text{hat}} - Y^{\text{bar}})$ for all i . However, negative deviations would cancel out positive deviations. The absolute values of the deviations could be used, but this poses particular problems when calculating other statistics based on the results. Hence, the most common method is to consider the sum of **squared deviations**. Thus we have:

$$\sum_{i=1}^n (Y_i - Y^{\text{bar}})^2 = \sum_{i=1}^n (Y_i - Y_i^{\text{hat}})^2 + \sum_{i=1}^n (Y_i^{\text{hat}} - Y^{\text{bar}})^2$$

Or as we had before,

$\sum_{i=1}^n (Y_i - Y^{\text{bar}})^2 = \text{TSS} =$ the sum across the whole sample (from $i = 1$ to $i = n$) of the squared deviations of: the observed values (Y_i) from the mean value (Y^{bar}).

$\sum_{i=1}^n (Y_i - Y_i^{\text{hat}})^2 = \text{RSS} =$ the sum across the whole sample (from $i = 1$ to $i = n$) of the squared deviations of: the observed values (Y_i) from the predicted values (Y_i^{hat}).

$$\sum_{i=1}^n (Y_i^{\text{hat}} - Y^{\text{bar}})^2 = \text{ESS} = \begin{array}{l} \text{the sum across the whole sample (from } i = 1 \text{ to } i = n) \text{ of} \\ \text{the squared deviations of: the predicted values } (Y_i^{\text{hat}}) \\ \text{from the mean value } (Y^{\text{bar}}). \end{array}$$

The **mean square** for each entry is the sum of squares divided by the degrees of freedom (*df*). If the regression assumptions are met, the ratio of the *mean square regression* to the *mean square residual* is distributed as an F-distribution, with k and $n - k - 1$ degrees of freedom (where k is the number of parameters being estimated). F serves to test how well the regression model fits the data. If the probability associated with F is small, the hypothesis that $R^2 = 0$ is rejected. With just one regressor, the square root of the calculated F value is equivalent to the t statistic for the slope coefficient.

A useful measure of overall goodness of fit of the regression line, and one that we have already come across, is the proportion of improvement due to the model:

$$\begin{aligned} R^2 &= \text{regression sum of squares} / \text{total sum of squares} \\ &= \text{REGSS} / \text{TSS} \\ &= \text{proportion of the variation of } y \text{ that can be explained} \\ &\quad \text{by the model} \end{aligned}$$

Remember from last week that R^2 is the proportion of the variation in y that is explained by the regression. Thus, the explained sum of squares is equal to R^2 times the total variation in y :

-
3. Comment on the Regression, Residual, and Total sum of squares from the output from the last two questions (i.e. whether relatively large or small, and the meaning of this).
 4. Confirm that $\text{TSS} = \text{REGSS} + \text{RSS}$ for both regressions.
 5. Confirm that $R^2 = \text{REGSS} / \text{TSS}$ for both regressions.
-

2.5 The F-Test

These sums of squares, particularly the RSS, are useful for doing hypothesis tests about groups of coefficients. The test statistic used in such tests is the F distribution:

$$F = \frac{(RSS_R - RSS_U) / r}{RSS_U / (n - k - 1)}$$

Where:

- RSS_U = unrestricted residual sum of squares (i.e. the RSS from the unrestricted regression)
 - = RSS under H_1 = RSS
- RSS_R = restricted residual sum of squares (i.e. the RSS from the restricted regression)
 - = RSS under H_0 = TSS
 - ($RSS_R =$ TSS under H_0 because if all coeffs were zero, the explained variation would be zero, and so error element would comprise 100% of the variation in TSS, I.e. RSS under $H_0 = 100\%$ TSS = TSS)
- r = number of restrictions
 - = number of slope coefficients in the regression that we are restricting
 - = equals all slope coefficients = k

For this particular test (i.e. when we are using F to test the hypothesis that all coefficients = 0), the F statistic reduces to $(R^2/k)/((1-R^2)/(n-k-1))$ so it isn't telling us much more than the R^2 . This formula is used in the ANOVA table produced by SPSS to test the null hypothesis that all slope coefficients equal zero. The hypothesis test has 4 steps:

- (1) $H_0: \beta_k = 0 \forall k$
 $H_1: \beta_k \neq 0 \forall k$
- (2) $\alpha = 0.05$,
- (3) Reject H_0 iff $\text{Prob}(F > F_c) < \alpha$
- (4) Calculate $P = \text{Prob}(F > F_c)$ and conclude
 (P is the "Sig." value reported by SPSS in the ANOVA table).

The F-test can also be thought of as the ratio of the mean regression sum of squares and the mean residual sum of squares:

$$F = \text{regression mean squares} / \text{residual mean squares}$$

If the line of best fit is good, F is large since the improvement in prediction due the regression will be large (so regression mean squares is large). As a result, the difference between the regression line and the observed data will be small (residual MS is small).

6. Calculate the F statistic for the two regressions you ran earlier using the general formula:

$$F = \frac{(RSS_R - RSS_U) / r}{RSS_U / (n - k - 1)}$$

7. Now calculate the F statistics using the $(R^2/k)/((1-R^2)/(n-k-1))$ formula.
8. Now calculate the F statistic using the ratio of mean squares formula:

$$F = \text{regression mean squares} / \text{residual mean squares}$$

2.6 Regression assumptions

For estimation of α and β and for regression inference to be correct we have to assume the following:

- i) *Equation is Correctly Specified*

That is, the equation is:

- Linear (Failure implies *functional form misspecification*)
- Contains all relevant variables (Failure implies *omitted variables*)
- Contains no irrelevant variables (Failure implies *inclusion of irrelevant variables*)
- Contains only variables that are measured without error (Failure implies *errors in variables*)

Failure to correctly specify the regression equation can make the estimates for α and β biased and inconsistent (estimates are consistent if as the sample size increases, the variance tends to zero, and the means of the repeated estimates tend towards the population values for α and β).

- ii) *Error Term has Zero Mean*

That is, the long run average (called the “*expected value*”) of the error term equals zero:

$$E(\varepsilon_i) = 0 \text{ for all } i.$$

(Failure implies non zero mean)

- iii) *Error Term has Constant Variance (Homoscedasticity)*

That is, $\text{Var}(\varepsilon_i) = \sigma^2$, where σ^2 is constant for all i .

(Failure implies Heteroscedasticity).

If Heteroscedasticity exists, the OLS estimates of α and β may be biased and inconsistent (but not necessarily so), and their standard errors (see below) will be wrong. This latter problem implies that the t and F statistics will be unreliable. Heteroscedasticity can be caused by a non-constant coefficient (eg $\beta_i = \beta + \varepsilon_i$), omitted variables, or the aggregation methods used for certain variables. This latter cause produces “*genuine heteroscedasticity*” which does not result in biased or inconsistent estimators.

There are a number of detection tests for heteroscedasticity (eg Glejser, Goldfield Quandt, Breusch Pagan), each one detecting a different form of heteroscedasticity. The G-Q (Goldfield Quandt) test, for example tests for heteroscedasticity of the form:

σ^2 is monotonically related to variable Z_i

First, all observations of all variables in the regression are ordered according to Z_i , and p central observations are removed. Separate regressions are then fitted to both lower (n_1) and upper (n_2) sets of observations. The test statistic G is then calculated from RSS_2 / RSS_1 (assuming σ^2_i is positively related to variable Z_i) which has an F distribution of the form $F [((n-p)/2-k), ((n-p)/2-k)]$. If G is greater than the critical value (chosen from the tables) for F , then the null hypothesis that there is no heteroscedasticity is rejected.

iv) *No Autocorrelation*

That is, if we are using time series data, $Cov(\epsilon_t, \epsilon_s) = 0$; where t does not equal s . In plain English this means that this years error term is not correlated with error terms from previous years.

v) *Explanatory Variables are Fixed Regressors*

That is, $X_1, X_2 \dots X_k$ are non stochastic (ie they are fixed in repeated samples as in an experiment). Failing this, we can do the regression analysis conditional on the regressors. This allows the regressors to be stochastic, but assumes that they are uncorrelated with the error term. If this assumption breaks down we have the problem of “simultaneity”, which implies that the estimates of α and β are not consistent.

vi) *Data Matrix has Full Rank*

That is, there is no linear dependence between RHS (right-hand side) variables. (Failure implies perfect multicollinearity which means that OLS is not possible. More common is imperfect multicollinearity, which results in estimates being unstable across sample sizes).

If the above assumptions are met, OLS estimates are said to be BLUE:

| | |
|-------------------|--|
| <u>B</u> est | I.e. most efficient = least variance |
| <u>L</u> inear | I.e. best amongst linear estimates |
| <u>U</u> nbiased | I.e. in repeated samples, mean of $b = \beta$ (where β is slope coefficient if the regression were run on the population and not just a sample = the “true” vale of the slope) |
| <u>E</u> stimates | I.e. estimates of the population parameters. |

9. try to memorise the assumptions underlying ordinary least squares (i.e. regression analysis)

10. Read chapters 1 and 2 of Kennedy “A Guide to Econometrics”. Pay particular attention what Kennedy has to say about the meaning and importance of “efficiency” and “bias”.

11. Read, Christopher H. Achen's *Interpreting and Using Regression* (London: Sage, 1982) and Chapter 4 of Andy Field, "*Discovering statistics using SPSS for Windows : advanced techniques for the beginner*".
-