1 Correlation and Inference from Regression

Reading:
Achen (1982) section 4 "Sampling Distributions of Regression Coefficients"
Kennedy (1998, 4th Ed.) "A Guide to Econometrics", Chapters 1 to 4
Field (2005, 2nd Ed.) Chapter 5. "Regression"

Aim:
The aim of this chapter is to introduce the notion of covariance and correlation as a means of ascertaining the relationship between variables and to introduce regression analysis.

Objectives:
By the end of this chapter, you should know how to compute the covariance and the correlation coefficients and understand their meaning. Readers should also able to understand the intuition behind regression estimation and the assumptions necessary for OLS to be BLUE. You should also be familiar with multiple regression and how to interpret coefficients, standard errors and R^2.

Plan:

1.1 Introduction

Often of greatest interest to social scientists is the relationships between variables, rather than just the nature of the variables themselves. We might want to know, for example, whether social class is related to political perspective; or whether there is a relationship between income and education; or whether worker alienation is related to job monotony. In this section we look at one of the simplest ways of gauging the strength of relationships between scale variables: the correlation coefficient. We then note the limitations of this approach and introduce multiple regression, which allows us to examine the relationship between several variables.

1.2 Covariance and the Simple Correlation Coefficient

The sign of the covariance between x and y will indicate the direction of the covariation of x and y, but its magnitude will depend on the scales of measurement (e.g. if y is miles per gallon, and x is average speed, then the cov(x,y) will be greater if one measures x in km per hour rather than miles per hour). One way round this scaling problem is to divide the covariance by the product of the standard deviations of x and y. This gives us the *Simple Correlation Coefficient*.

To compute correlation coefficients in SPSS, go to Analyse, Correlate, Bivariate and choose which variables you wish to analyse. Then select Pearson's Correlation Coefficient, and choose whether you want two tailed or one tailed tests of significance. The null hypothesis in these tests is that there is no correlation between the variables. H_0 can be rejected if the significance level is small (e.g. < 0.05). Note that the correlation coefficient only measures *linear* relationships between variables. Click OK to obtain the table of results or click Paste to retain the syntax.

1.2.1 Exercise:

1. (i) Load the Nationwideextract data set.

   (ii) Create a variable that measures the age of the dwelling in 1999 (e.g. age_dw = 1999 – dtbuilt).

   (iii) Now run Bivariate Correlations between purchase price, floor area and age of dwelling.

   (iv) What evidence is there that: (a) price and floor area are correlated; (b) price and age of dwelling are correlated; (c) age of dwelling and floor area are correlated?

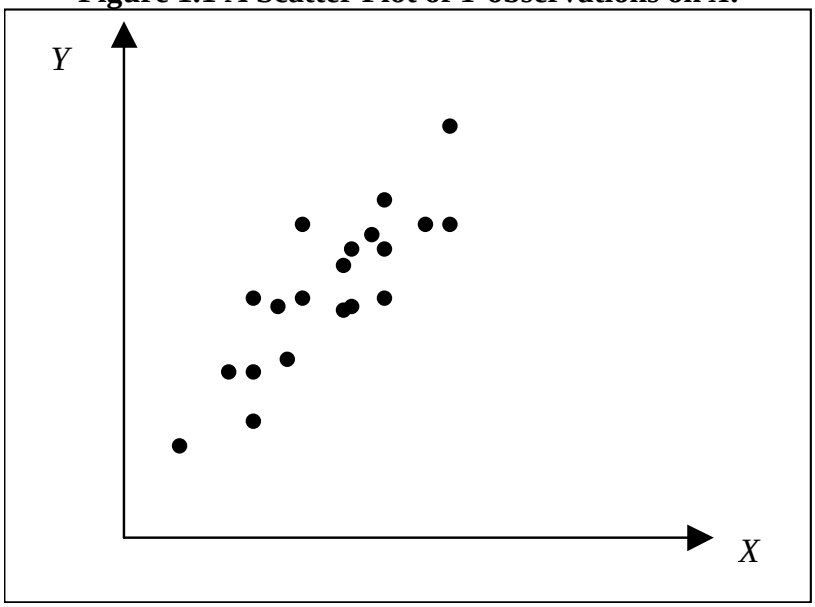
2. Create a variable equal to the square of age of dwelling. Is there any correlation between price and the square of age of dwelling? Compare you result to 1(b) above.

3. Run a correlation matrix for number of bedrooms, floor area, and number of bathrooms. Comment on your results.

1.3 Ordinary Least Squares And Related Terminology

Consider two variables, X and Y. You believe Y to be determined by X; that is, you hypothesise that X is an "independent" variable (ie not determined by Y), and Y is a "dependent" variable (ie dependent on X). Now imagine a scatter plot of a sample of Y observations, such as in Figure 1, for example. We can see from the scatter of points that there may be some relationship between the two variables, possibly a linear (i.e. "straight line") relationship. We could guestimate the nature of this relationship by using a ruler to draw what looks to be the line of best fit.

Figure 1.1 A Scatter Plot of *Y* observations on *X*.



Alternatively, we could use a simple mathematical formula called OLS (ordinary least squares). This draws the line that minimises the sum of squared deviations from the line to each scatter point above and below it. As with any linear relationship, the equation of the computed line will comprise of a *Y* intercept (also called "the constant") denoted by α, and a slop coefficient donated by β:

$$Y = \alpha + \beta X$$

If this is the "true" or "underlying" relationship between *Y* and *X* that we would observe from the population, we use OLS to *estimate* values for α and β (usually denoted α-hat and β-hat respectively). However, because any straight line we draw through the scatter points will rarely fit exactly (i.e. deviations will not be zero), even when run on the population, we have to include an error term, ε. Thus:

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where *i* is the i^{th} observation of *Y* and *X*. Note that once we have estimated the straight line relationship between Y and X, we can that estimated relationship to tell us what we would predict the value of Y to be for any value of X. So, suppose we only have 80 observations on X and Y and we want to estimate what the value of Y would be if X was 23.7. Of the 80 observations on *X*, none of them equal 23.7. Now, once we have estimated the values of α and β, we can use them to give us a predicted value for Y given that X = 23.7:

$$Y_i = \alpha + \beta \times 23.7 + \varepsilon_i$$

Ignoring the error term for now, let's assume that our estimates of α and β, come out as 3.6 and 7.4 respectively, then the value of *Y* when *X* is 23.7 is calculated as follows:

$$Y_i = 3.6 + 7.4 \times 23.7$$
$$= \underline{\mathbf{178.98}}$$

1.4 OLS Estimates are BLUE

One of the major attractions of using OLS is that given certain assumptions the OLS estimates for α and β (α-hat and β-hat) are **BLUE** (**B**est **L**inear **U**nbiased **E**stimators). They are "**Best**" in that they have the minimum variance compared with other estimators (i.e. given repeated samples, the OLS estimates for α and β vary less between samples than any other sample estimates for α and β). They are "**Linear**" in that a straight line relationship is assumed. They are "**Unbiased**" because, in repeated samples, the mean of all the estimates achieved will tend towards the population values for α and β. And they are "**Estimates**" in that the true values of α and β cannot be known, and so we are using statistical techniques to arrive at the best possible assessment of their values, given the information available.

For estimates to be BLUE, however, certain assumptions have to be met. These assumptions, and what happens when they fail, and considered in section 9.2 below. First, however, it should be noted that so far we have only considered the case of one RHS (right-hand side) variable. This is termed *Simple Regression*. When we have more than one explanatory variable, (eg $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$) the analysis becomes that of Multiple Regression. Usually *X*s are assumed to be variables that influence *Y* (*Y* is said to be dependent on the *X*s). It should also be noted that there are several alternative terms used in the literature (and in the report) for *Y* and $X_1, X_2 \ldots X_k$. This is often confusing to someone new to the subject, so I have reproduced the following table from Maddada (1992, p.61).

Each of the terms in **Table 1** is relevant for a particular view of the use of regression analysis. Terminology (a) is used if purpose is prediction. For instance, sales is the predictand and advertising expenditures is the predictor. The terminology in (b), (c) and (d) is used by different people in their discussion of regression models and are largely equivalent. Terminology (e) is used in studies of causation. Terminology (f) is specific to econometrics and should be used with great care following Engle, Hendry and Richard (1983, "Exogeneity", *Econometrica*, Vol 51, March). They redefine exogeneity in terms of weak, strong and super exogeneity. Finally, terminology (g) is used in control problems. For instance, our objective might be to achieve a certain level of sales (target variable) and we would like to determine the level of advertising expenditures (control variable) to achieve out objective.

Table 1 Classification of Variables in Regression Analysis

Y	$X_1, X_2 \ldots X_k$
(a) Predictand	Predictors
(b) Regressand	Regressors
(c) Explained variable	Explanatory variables
(d) Dependent variable	Independent variables
(e) Effect variable	Causal variables
(f) Endogenous variable	Exogenous variables
(g) Target variable	Control variables

1.5 Doing Regression analysis in SPSS

To run regression analysis in SPSS, click on Analyse, Regression, Linear. You will then need to select your dependent (i.e. 'explained') variable and independent (i.e. 'explanatory') variables.

If you click OK once you've entered your variables, you will end up with a series of output tables including a **Model Summary** (details on the overall goodness of fit as measured by R square), **ANOVA** (Analysis of Variance), and **Coefficients** (the estimates of the slope and intercept terms).

1.5.1 Example

Suppose we believe that there is a relationship between number of bathrooms and the floor area of a house. One might theorise that the more bathrooms there are, the larger the floor area, given that bathrooms take up space. So our theory suggests the following relationship (assumed to be linear) between area and bathrooms:

$$\text{Floor area} = \alpha + \beta \,\text{Number of bathrooms} + \varepsilon$$

If we run this regression in SPSS using the Nationwideextract data, we would get the following output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.584[a]	.341	.340	35.58

a. Predictors: (Constant), Number of Bathrooms

The "Model Summary" table includes the R square which tells you the proportion of the dependent variable that your independent variables explain (in this case = 0.341 = 34.1%). If you have **more than one explanatory variable** you should use the **adjusted R square** which controls for the fact that the R-square will rise each time you add an additional variable even if there is no real gain in the explanatory power of the model.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	362375.7	1	362375.7	286.292	.000[a]
	Residual	701228.5	554	1265.755		
	Total	1063604	555			

a. Predictors: (Constant), Number of Bathrooms

b. Dependent Variable: Floor Area (sq meters)

The most useful information in the ANOVA table at this stage is the F statistic. This tests the H_0 that all slope coefficients are jointly equal to zero. It is another summary test of the whole model and is

related to the R square measure. If "Sig" is small, you can confidently reject the null. Also useful is the total degrees of freedom which tells you at a glance how many observations were included in the model (in this case 555+1).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.928	4.700		8.708	.000
	Number of Bathrooms	64.622	3.819	.584	16.920	.000

a. Dependent Variable: Floor Area (sq meters)

The Coefficients table gives you the estimate of the intercept (Constant) and slope coefficient under column "B". The t statistic tests the hypothesis that B = zero and is calculated by dividing B by the Standard Error of B. If Sig. is small you can confidently reject the null of B = 0.

The Coefficients table tells us that floor area of a house in the Nationwideextract sample increases by 64.6m^2 with every extra bathroom, and that for a house with no bathrooms, the total floor area of the house would be 40.9 m^2. Because the t value is large for the slope coefficient (resulting in a small significance level = 0.000), we can reject the null hypothesis that the slope coefficient = 0 (i.e. that there is no relationship between floor area and number of bathrooms).

Similarly, the *t* value is large for the intercept term, and so we can reject the null hypothesis that the slope coefficient is zero.

1.6 Multiple Regression

The problem with both correlation and simple regression is that neither allow you to control for the effect of other variables (e.g. there may be a strong simple correlation between income and education, but if one controls for IQ, then there may be a much smaller correlation between education and income). One way of overcoming this is to use multiple regression. Multiple regression is regression analysis when you have more than one explanatory variable.

To do regression analysis in SPSS, go to Analyse, Regression, Linear. Decide on your Dependent and Independent variables from the list on the left hand side. If you then click OK (or Paste and run the syntax) SPSS will present you with the results of the OLS regression.

1.7 Interpreting coefficients

In a simple regression (one explanatory variable), the regression estimates the line of best fit: ie it estimates the values of the intercept α and slope β:

$y = a + bx$ where a and b are sample estimates of the population parameters α and β

When you have two variables, the regression estimates the *plane* of best fit (that is, a surface in three dimensions that best fits the 3D data points):

$$y = a + b_1x_1 + b_2x_2$$

(Beyond two explanatory variables, there is no straighforward graphical representation of the fitted surface).

The slope coefficients represent the amount the dependent variable would increase for every additional unit of the explanatory variable. In the income/education/experience example given in the lecture, the estimated equation of the plane was:

$$y = -4.20 + 1.45x_1 + 2.63x_2$$

so for every extra year of education (x_1) income rises by £1,450 and for every extra year of experience (x_2) income rises by £2,630.

1.7.1 Exercise:

1. Using data from Nationwideextract:

   (i) Do a scatter plot of the relationship between purchase price and floor area.

   (ii) Comment on the plot and then insert an OLS line of best fit onto the plot. (To add a line of best fit, double click on the graph, right-click on the data points on the chart, and select "Add Fit Line at Total" then select Linear, and click Close. Then click File, Close to close the graph editor window and return to the Output Viewer).

   (iii) How well do you think the regression line fits the data?

   (iv) Run a linear regression to obtain the numerical values of the relationship.

   (v) What does the statistical output from the coefficients table tell us about the relationship?

2. (i) Run a 3-D Scatter plot with number of bedrooms as the second explanatory variable. It sometimes helps to see where there data points are in space if you include "spikes" either to the origin or to the floor. To do this, double click on the graph and right click on the observations. Then select Properties from the drop-down list, click the Spikes tab and choose the type of spikes you want. To rotate the graph, double click on it and then right-click on the observations and select 3-D Rotation.

   (ii) Comment on the graph.

(iii) Run a regression equation with the second explanatory variable, bedrooms, included along with floor area.

(iv) Has the inclusion of the extra variable added anything to the explanatory power of the model?

(v) Try replacing it with number of bathrooms and comment on your results.

3. (i) Experiment with a number of 2-explanatory variable models, comparing the 3-D scatter plot with the regression output.

(ii) Then experiment with more than 2 explanatory variables.

Further Practice:

4. Run a multiple regression of purchase price on floor area and age of dwelling. What do the estimated coefficients mean?

5. Now double click on the regression tables and right click on each term and select "What's this?". This will give you brief explanations to help you interpret the rest of the table. (This facility is available for most SPSS output tables).

6. Change the label on the "terrace" variable so that it reads "Terraced House", and give the variable new labels: = 1 if terraced and = 0 otherwise. Do a 3-D scatter plot of purchase price, floor area and terrace. What sign would you expect the coefficient on terrace to have if it were included in a regression of price, floor area and terrace? Run this multiple regression to check whether your intuition is accurate. What does the value of the coefficient on terrace mean?

1.8 Inference from regression

The estimates of the slopes and intercept are however subject to error since we are usually looking at a sample of observations, not the population. The standard deviations of a, b_1 and b_2 are called *standard errors* since a, b_1 and b_2 are long run averages (*expected values*) of what you'd get if you run the regression on lots of different samples from the same population in much the same way that the mean is the expected value of the population a, b_1 and b_2 are the sample estimates of the slopes and intercept that you'd get from running a regression on the population the range of values you'd get for a parameter from repeated samples is called the *sampling distribution*. The standard error reported in SPSS for a particular coefficient is an estimate of the standard deviation of the sampling distributions of the coefficient.

1.8.1 Confidence Intervals

We can use the standard error of the estimated coefficient to calculate the confidence interval for the population parameter β_i:

$$\beta_i = b_i \pm t_c \mathrm{SE}(b_i) \text{ with } df = n - k$$

where

k = number of coefficients being estimated including the constant

= 1 + no. of variables in the regression.

The value of t_c will depend on what level of confidence we want and the *df*. For example:
at 95% level of confidence & $df = 2$, $t_c = 4.303$
at 80% level of confidence & $df = 2$, $t_c = 1.886$

1.8.2 Exercise

1. Use the formula above to calculate the 95% confidence interval for the coefficients in the output from the last question (i.e. the coefficients for intercept, floor area and terrace). Note that because the degrees of freedom are large ($df = n - k = 552$), the critical t-value for the 95% confidence interval will be 1.960 – you should verify this either using SPSS or t-tables).

2. Now use SPSS to calculate the confidence intervals: go to Analyse, Regression, Linear…, Statistics, and check the Confidence Intervals option. Then click Continue and run the regression. You will get a repeat of your regression results but with an extended table including confidence intervals for the coefficients. How do the results compare with your own calculations of the confidence intervals? (there may be a slight difference due to rounding).

1.8.3 Hypothesis tests on β_i

The t-values provided in the SPSS output test the null hypothesis that $\beta_i = 0$ (i.e. that there is no relationship between y and x_i). The t value is calculated by dividing the coefficient by its standard error and has n-k degrees of freedom. For samples greater than 120, the critical t-value at the 5% significance level is 1.96. So, as a rule of thumb, we say that an explanatory variable is statistically significant (i.e. its coefficient is significantly different from zero) if the t value is greater than 2. As you would expect, the larger the t-value the smaller the probability of a type I error (i.e. the probability of rejecting H_0 when it is in fact correct). The probability of a type I error is given in the column after the t-value, labelled "Sig." (if this is < 0.05, you can reject the null at the 5% level of significance – i.e. if you reject the null, there is only a one in twenty chance that you have done so incorrectly).

1.8.4 Exercise

1. Run a regression of purchase price against floor area, number of bedrooms, number of bathrooms, age of dwelling, and terrace. What do the t-values against each of these variables tell you? What do you think you should do with variables that have low t-values (i.e. large Sig. values)?

1.9 Partial Correlation Coefficients and the coefficient of determination

One of the advantages of multiple regression is that it allows us to calculate the partial correlation coefficient: the correlation between y and x_2 controlling for the effect of x_1. We shall come back to these when we look at multicolinearity.

One useful measure that is worth looking at this stage is the Coefficient of Multiple Determination, or $\mathbf{R^2}$. This measures the proportion of the variation in y explained by all the explanatory variables and is a good measure of the overall goodness of fit of the regression line or surface. It varies between 0 and 1; the nearer it is to 1, the more of y that is being explained and so the better the goodness of fit.

Each time an explanatory variable is added to the regression, the R^2 will rise even there is no real gain in explanatory power. Thus, where more than explanatory variable is included you need to look at the **Adjusted $\mathbf{R^2}$** which attempts to correct for this.

1.9.1 Exercise

1. Look at the Adjusted R^2 in each of the various regression outputs you have produced so far. Which model of house price determination appears to be the best on this basis?
