CA200 – Quantitative Analysis for Business Decisions

File name: CA200_Section_06_RegressionIntro

Table of Contents

6. Introduction to Regression and Correlation

6.1 Overview

Linear Regression:

In Statistics, **regression analysis** includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a **dependent variable** (***most often called y***) and one or more **independent variables**. We will deal with the case of one independent variable only (***most often called x***). The aim is to understand how a typical value of the dependent variable changes when the independent variable is varied.

We will deal only with **Linear Regression**, where there is assumed to be a *straight line relationship* between dependent and independent variables.

The input to the regression process is a sample of pairs of values of x and y.

Outline of regression /straight line 'fitting'

It is always a good idea to create a **Scatter Plot** of the data beforehand as a visual check that the assumption of a straight relationship is *plausible*.

Be clear on which variable is *independent* and which is *dependent*. The regression of y on x, (y *dependent*, x *independent*) is *NOT* the same as the regression of x on y (x dependent, y *independent*).

We will use the notation $y = \alpha_R + \beta x$ for the linear regression straight line equation.

Like any straight line it is defined by two parameters; here α_R is the value of y when x = 0 (called the **intercept** on the y axis) and β is the **slope** of the line.

Note 1: the use of the R subscript for alpha (the intercept) is just to remind you that this refers to the regression line and is not the same thing as the *level of significance* or *risk* for hypothesis testing, which defines the size of the rejection region and is conventionally labelled α.

We can consider the regression process as a means of summarising all the input data by just two values. This is similar to how one summarises a set of values by one number, their average.

Note 2: The implication of this is that we now have an **average line**, rather than a **single average value**. Each point on the line acts as a mean value for the range of possible y values at that particular x.
(This is why regression is sometimes called a *many-sample technique*, even though we only ever deal with the single 'sample of paired x and y values' and *one* mean line with two parameters in simple linear regression).

Correlation Coefficient (r):

A correlation coefficient (**r**) is a single number (an *index*) that describes the *degree of association* between two variables. Its range is $-1 \leq r \leq +1$. If **r = -1**, variables are perfectly **negatively** correlated, if **r = 0** there is **no correlation** and if **r = +1** the variables are perfectly **positively** correlated.
It is very often computed at the same time as the regression line, but is **not** the same thing as the slope, **nor** does it require knowledge of which (if either) variable is *dependent* or *independent*, as it is an *association* measure. A general expression is thus (see statistical tables):

$$r = \frac{COVAR(x, y)}{\sqrt{(VAR(x)(VAR(y)}}$$

Of interest also is $\mathbf{r}^2$ as we can interpret it as a measure of how well the independent variable *x* explains the variation in *y*, i.e. how closely the independent and dependent variables are related. This quantity is called the **coefficient of determination**.

6.2 Formulae for α_R, β and r

We start with a sample of *n* pairs of observations: $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ ….. (x_n, y_n)
We want to:

Obtain the relationship between *y* and *x*

Estimate *y* for a particular value of *x*:

Note:

It is also possible of course to give formulae for **confidence intervals** for α_R and β (or for the set-up of the complementary **hypothesis test**).

These are similar to what we had before for obtaining a confidence interval (or hypothesis test) for the mean, for which we needed to know the sample values, the standard error of the mean and the appropriate distribution, as well as the level of risk we wanted to take.

As we now have a **mean line**, the S.E. needs a bit more work to calculate, so we will outline only what is involved in setting up a simple hypothesis test.

The "true" regression line is represented by the equation $y = \alpha_R + \beta x$ where the unknowns α_R and β are to be ***estimated*** from the sample data. These sample estimates are denoted a and b. So, the sample regression line is $y = a + bx$

Previously, we estimated the unknown mean μ by calculating the sample value and establishing the confidence intervals or test basis for the population parameter.

In fact, if $\beta = 0$ there would be *no dependence* of y on x and α_R would just be the same as μ with "a" the same as $\bar{x}$, i.e the data could be summarised by a single mean value.

Estimation: Parameters α_R and β **for the regression line and of *r* for correlation.**

The principle is to

Minimise the sum of the squared distances (S) from the sample points to the line.

This means we choose ***a*** and ***b*** to minimise

$$S = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

This is the basis for the description ***least squares estimation***.

We do not present the derivation but the resulting formulae are:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$a = \bar{y} - b\bar{x}$$

Finally, the formula for the correlation coefficient '**r**' , using the variance and covariance information is:

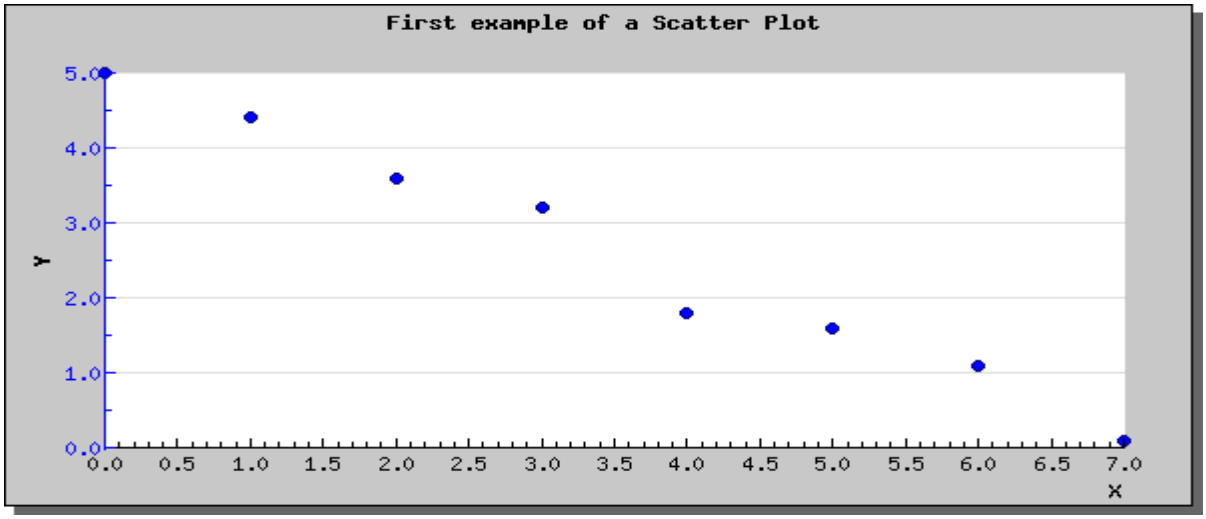
$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2)(n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}}$$

6.3 First example using the formulae for α, β and r

Calculate the least squares regression line and the correlation coefficient for the following set of data. First display the data on a scatter plot.

x	0	1	2	3	4	5	6	7
y	5	4.4	3.6	3.2	1.8	1.6	1.1	0.1

Solution



So we can see that there is indeed a very plausible *linear relationship*. Y decreases as x increases, so it appears to be a negative relationship also.

Calculation Procedure:

It is best to use a tabular lay-out as follows.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**SUMS**	**Means**
x_i	0	1	2	3	4	5	6	7	28	3.5
y_i	5	4.4	3.6	3.2	1.8	1.6	1.1	0.1	20.8	2.6
x_i^2	0	1	4	9	16	25	36	49	140	----
$x_i\, y_i$	0	4.4	7.2	9.6	7.2	8.0	6.6	0.7	43.7	----
y_i^2	25	19.36	12.96	10.24	3.24	2.56	1.21	0.01	74.58	----

Then slope,

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{8(43.7) - (28)(20.8)}{8(140) - (28)(28)} = -0.69$$

and intercept

$$a = \bar{y} - b\bar{x} = 2.6 + 0.69 * 3.5 = 5.03$$

So, the equation of the 'fitted' straight line is:

$$y = 5.03 - 0.69x$$

While the correlation coefficient is:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2)(n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}} = \frac{8(43.7) - (28)(20.8)}{\sqrt{(8(140) - (28)(28)) * (8(74.58) - (20.8)(20.8)}} = -0.99$$

Which implies a *strong degree of negative association* – as indicated by the plot.

Further, we note that $r^2 = 0.98$, which means that 98% of the *variation in y* is accounted for by a *linear relationship* with *x*.

Hence, $r^2 = 0.98$ which means that 98% of the variation in *y* is accounted for by a linear relationship with *x*.

6.4 Second example using the formulae for α_R, β, r and r^2

For the following data have, gathered for a sample of 12 students,

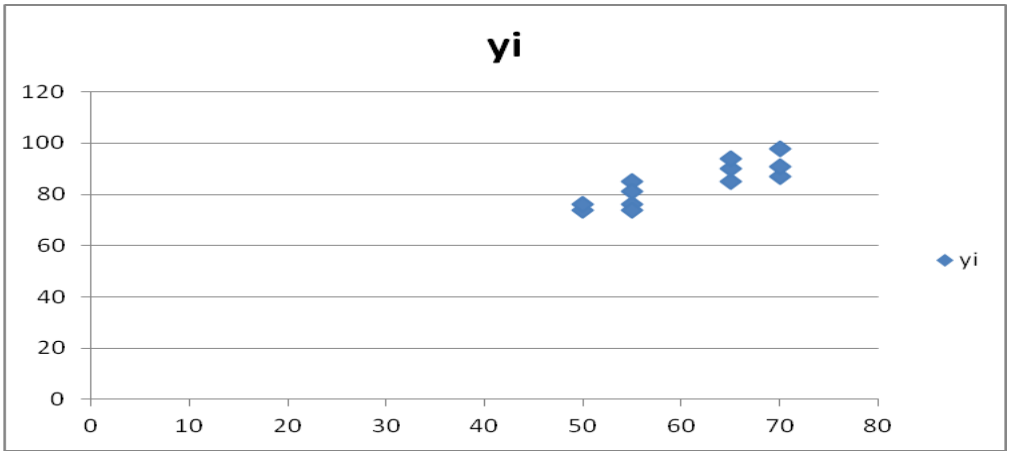
Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Last year's maths score (x_i)	65	50	55	65	55	70	65	70	55	70	50	55
This year's Stats score (y_i)	85	74	76	90	85	87	94	98	81	91	76	74

use these as an exercise to :

1. Draw a scatter plot to check the plausibility of there being a linear relationship between the two sets of scores.
2. Estimate the relationship between performance in maths and statistics
3. Predict the statistics result for a student who obtained a score of 60 in maths
4. Measure the degree of association between maths and statistics and the explained amount of variability in y due to x, (**coefficient of determination**).

Solution:

(1)



So we can see that there is a *plausible* linear relationship.

<u>Calculations</u>:

Again, these should be tabulated for convenience, as it makes for much easier checking of working.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**SUMS**	**Means**
$\mathbf{x_i}$	65	50	55	65	55	70	65	70	55	70	50	55	725	90.625
$\mathbf{y_i}$	85	74	76	90	85	87	94	98	81	91	76	74	1011	126.38
$\mathbf{x_i^2}$	4225	2500	3025	4225	3025	4900	4225	4900	3025	4900	2500	3025	44475	----
$\mathbf{x_i\, y_i}$	5525	1	4180	5850	4675	6090	6110	6860	4455	6370	3800	4070	57986	----
$\mathbf{y_i^2}$	7225	5476	5776	8100	7225	7569	8836	9604	6561	8281	5776	5476	85905	----

(2) Then, using the formulae given, should be able to check that b = 0.9 and a = 30.0.

So, Estimated relationship is

"This year's Stats score"$_{\text{STUDENT}}$ = 30.0 + 0.9*["Last year's maths score"$_{\text{STUDENT}}$]

(3) Hence, can show that the predicted Stats mark = 30.0 + 0.9*60 = 84

(4) Degree of association – can also show r = 0.86. Hence, r^2 = 0.74 which means that 74% of the variation in *y* is accounted for by a linear relationship with *x*

6.5 Estimation and Hypothesis Testing in Regression

- Estimates **a** and **b** are single (one-off) calculations from a set of sample data, (i.e. similar to the estimate of the mean or proportion from a single sample), so can regard as the *point estimate solution for the line*.
- Usually, therefore, we go on to provide an **interval estimate**, (i.e. whether sample results are likely 95% (or 99% or..) of the time to represent what is happening in the population) *or* we use the sample results as a basis for a **hypothesis test (**i.e. to make a statement about what we expect to be happening in the population, providing evidence from the sample **for** or **against** this hypothesis).
- We can also do this additional estimation and/or testing for the regression line, but it is a bit more work, so we focus here on just *outlining* the hypothesis test approach. (Both estimation and hypothesis testing use the same information as we have seen).

The information that we need includes the following elements:

- **Point estimate** – in this case as it is a line, we need both **a** and **b**
- **Hypotheses**: typically test if the intercept and, (independently), the slope are *zero*.
- **Distribution:** Use the t-distribution (conservative so works for *small* samples too)
- **Standard Error**: this involves calculating variances along the line –this is the part that involves most of the work
- **Level of significance** (α), i.e. level of *risk* that prepared to accept = *size of rejection region*
- **Decision Rule:** Decision again based on dividing up the distribution and seeing if the value of the Test Statistic generated by the sample data falls into the acceptance or rejection regions for the null hypothesis H_0

So Hypothesis Tests are of form:	$H_0 : \alpha_R = 0$	and independently	$H_0 : \beta = 0$
	$H_1 : \alpha_R \neq 0$		$H_1 : \beta \neq 0$
	Intercept		Slope

The first hypothesis set tests whether the line goes though the origin, where the null hypothesis says that it is expected to. The second set tests whether the slope is different from zero, where again the claim is that no dependence of y on x is involved. If the slope *is* effectively zero then this is in agreement with what the *null hypothesis* states, i.e. no straight line regression effectively.

Tests are of usual form: $T_{n-2} = \dfrac{Observedvalue - Expected\ value}{S.E.}$

So for the slope $T_{n-2} = \dfrac{b-0}{s\sqrt{\left(\dfrac{1}{\sum(x_i - \bar{x})^2}\right)}}$,

and for the intercept: $T_{n-2} = \dfrac{a-0}{s\sqrt{\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)}}$

where s (the standard deviation) is based on calculating all the deviations of the actual y values about the (*average*) line , squaring these, dividing by the degrees of freedom (= n-2 here) to give the variance, and then taking the square root to give s.

Decision rule

Again we divide up the distribution, (which is the t-distribution with n – 2 degrees of freedom here), according to the level of risk we are prepared to take, (i.e. the *level of significance*). The acceptance or rejection of the null hypothesis for *slope* (and similarly for *intercept*) then depends on whether the test statistic value, based on the sample data, falls into the *acceptance* or *rejection* regions of the distribution, as usual.