Least Squares

Introduction

We have mentioned that one should not always conclude that because two variables are correlated that one variable is causing the other to behave a certain way. However, sometimes this is the case for example in the example of Bumblebees it is the presence of nectar that attracts the Bumblebees.

In this section we will deal with datasets which are correlated and in which one variable, x, is classed as an **independent variable** and the other variable, y, is called a **dependent variable** as the value of y depends on x. As in the case of the nectar, x, and the bumblebees, y.

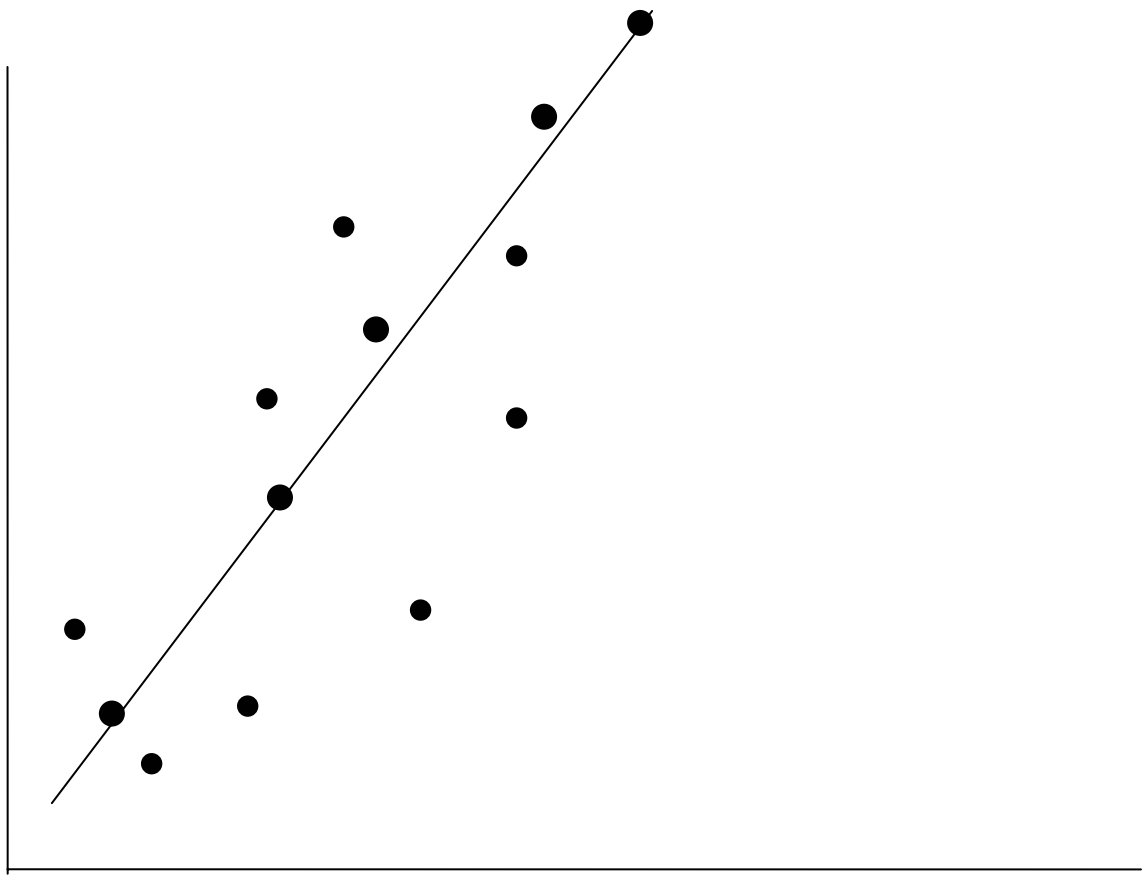
We saw that correlation implies a linear relationship. Well a line is described by the equation

$$\mathbf{y = a + bx}$$

where **b** is the **slope** of the line and **a** is the **intercept** i.e. where the line cuts the **y** axis.

Suppose we have a dataset which is strongly correlated and so exhibits a linear relationship, how

would we draw a line through this data so that it fits all points best?
We use the principle of least squares, we draw a line through the dataset so that the sum of the squares of the deviations of all the points from the line is minimised.



EXTRA EXPLANATION:

For each point in the dataset:

y - (a + bx) measures the vertical deviation (vertical distance) from the point to the line. Some points are above the line and **y - (a + bx)** will be positive for these points. For points below the line **y - (a + bx)** will be negative, so we square these deviations to make them all positive.

Now if we calculate **[y - (a + bx)]^2** for each point **(x,y)** and add them all up we get the sum of the squared distances of all the points from the line. The line which minimises this sum of squared distances is the line which fits the data best and we call it the **Least Squares Line**.

Using some calculus we can find the equation of this Least Squares Line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The Simple Linear Regression Model

If there is a linear connection between x and y in the population the model will be as below. We find that for a particular value of **x**, when an observation of **y** is made we get:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where **ε** (epsilon) is a random error component which measures how far above or below the True Regression Line (i.e. the line of means) the actual observation of **y** lies. The mean of **ε** is zero.

This model is called a Probabilistic Model because **ε** is a random variable.

Fitting the Model

The Simple Linear Regression Model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

contains 3 unknown parameters;

$\boldsymbol{\beta_0}$ - the intercept of the line,

$\boldsymbol{\beta_1}$ - the slope of the line

and $\boldsymbol{\sigma^2}$ the variance of $\boldsymbol{\varepsilon}$**.**

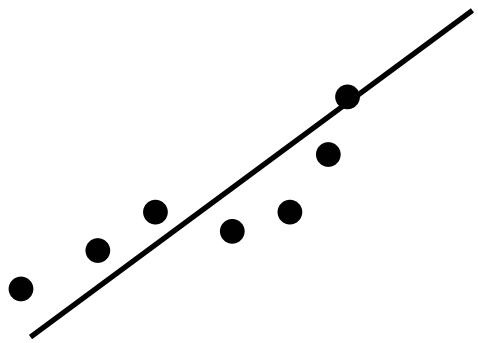
We will need to estimate these parameters (or population characteristics) using the data in our sample.

Remember in the past how we estimated the population mean μ using the sample mean $\bar{x}$ and the population standard deviation σ by the sample standard deviation s. This procedure is basically the same, we can in fact find Confidence Intervals for $\boldsymbol{\beta_1}$ and we can conduct a Hypothesis test for $\boldsymbol{\beta_1}$**.**

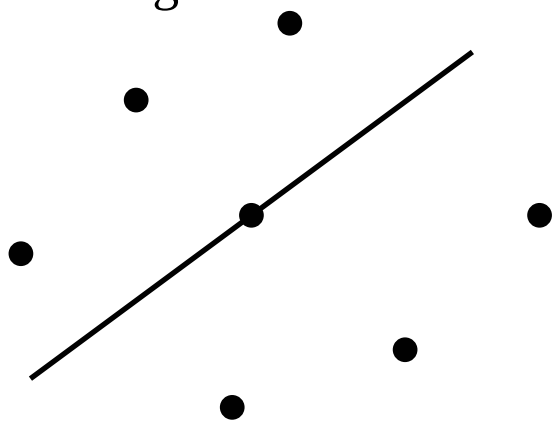
In fact we have already seen the Sample Statistics that we will use to estimate the Parameters (Population Characteristics) $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ they are $\hat{\beta}_0$ and $\hat{\beta}_1$ the intercept and slope of the least squares line (which are given in standard computer output).

Remember σ^2 measures how spread out the points are from the True Regression Line.

Small σ^2



Large σ^2



IMPORTANT: The Least Squares Line is an estimate, based on the sample, for the True Regression Line.

OK SO WHAT!

Well the true connection between any **y** and **x** is described by the probabilistic model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

and so even if we knew **x** and the coefficients $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ we couldn't determine **y** because of the random factor $\boldsymbol{\varepsilon}$.

Instead we use our sample data to find estimates for the coefficients $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ **ie:** $\hat{\beta}_0$ and $\hat{\beta}_1$.

We can then predict what the value of **y** should be corresponding to a particular value for **x** by using the Least Squares Prediction Equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad .$$

Where $\hat{y}$ is our prediction for **y**.

The Coefficient of Determination

Now that we have a model to describe the relationship between x and y we might, after fitting the data to our Least Squares Line, ask ourselves the question:

"How good is this Model at explaining the behaviour of the y values?"

EXTRA EXPLANATION:

According to our Linear Regression Model most of the variation in y is caused by its relationship with x. Except in the case where all the points lie exactly on a straight line (ie where r = +1 or r = -1) the model does not explain all the variation in y. The amount that is left unexplained by the model is SSE.

Suppose that the variation in y was not caused by a relationship between x and y, then the best estimate for y would be $\bar{y}$ the sample mean. And the Sum of Squared Deviations of the actual y's from this prediction $\bar{y}$ would be $SS_{yy} = \sum (y - \bar{y})^2$.

If little or none of the variation in y is explained by the contribution of x then SS_{yy} will be almost equal to SSE. If all of the variation in y is explained by its relationship with x then SSE will be zero.

Definition:
The coefficient of determination is

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

This represents the proportion of the total sample variability in y that is explained by a linear relationship between x and y.

IMPORTANT:
R-Squared measures how well the model fits the data. Values of R-Squared close to 1 fit well. Values of R-Squared close to 0 fit badly.

The Model Utility Test

There is one specific Hypothesis test that has a special significance here.

The test $H_0: \beta_1 = 0$

Vs

$H_A: \beta_1 \neq 0$

tests whether the slope of the regression line is non-zero.

Why is this special?

If y really depends on x then x should be a term in the final model. If $\beta_1 = 0$ then that is equivalent to saying that x is not in the model. So the variation in y is random and not dependent on x.

The Hypothesis test $H_0: \beta_1 = 0$ Vs $H_A: \beta_1 \neq 0$ therefore tests whether the model is **USEFUL**. If we **reject** the Null hypothesis for this test then we conclude that $\beta_1 \neq 0$ and so x should be in the model ie: the Model is **USEFUL**.

Hence the name Model Utility Test.

More complicated models

y depends on several variables x_1 , x_2 ,....,x_n

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

How can we tell if this model is useful?

Calculate r^2?

Or do a model utility test?

For which variables?

A political scientist wants to use regression analysis to build a model for support for Fianna Fail. Two variables considered as possibly effecting support for Fianna Fail are whether one is middle class or whether one is a farmer. These variables are described below:

y = Fianna Fail Support
x_1= Proportion Farmers
x_2= Proportion Middle Class

40 samples were chosen and opinion polls conducted on these samples to determine 40 sets of values for the three variables. The model is chosen to be :

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 .$$

A Computer printout for the interaction model is shown below.

PREDICTOR VARIABLES	**COEFFICIENT**	**STD ERROR**	**STUDENT'S T**	**P-value**
CONSTANT	0.0897	14.1270	0.00	0.9959
Middle Class	12.2	4.8	2.54	0.0479
Farmers	4.4	44.30	0.15	0.9234
Middle Class-Farmers	5.9	1.55	3.56	0.0164

R-SQUARED 0.8043
ADJUSTED R-SOUARED 0.7595

RESID. MEAN SQUARE (MSE} 155.325
STANDARD DEVIATION 12.5001

SOURCE	**DF**	**SS**	**MS**	**F**	**p -value**
REGRESSION	3	457.8	152.565	7.6	0.040
RESIDUAL	36	700	19.416		
TOTAL	39	1157.8			

CASES INCLUDED 12

MISSING CASES O

Interpreting this Computer output.

First look for R-Squared or better still Adjusted R-Squared. As mentioned before we want values close to 1.

Secondly look at the p-value in this table:

SOURCE	DF	SS	MS	F	p –value
REGRESSION	3	457.8	152.565	7.6	0.040
RESIDUAL	36	700	19.416		
TOTAL	39	1157.8			

If this is smaller than or significance level. Then the model is useful as a whole.

Now we can examine the individual terms in the model. For this part we look at the p-values in this table:

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P-value
CONSTANT	0.0897	14.1270	0.00	0.9959
Middle Class	12.2	4.8	2.54	0.0479
Farmers	4.4	44.30	0.15	0.9234
Middle Class-Farmers	5.9	1.55	3.56	0.0164

Again p-values smaller than our chosen significance level (usually 0.05) indicate variables that should be in our final model. P-values larger than our significance level may be left out of the model.