MM104 / MM107
Statistics and Data Presentation Semester 2

Project 6: Linear Regression

Ainsley Miller
ainsley.miller@strath.ac.uk

Lecture Overview

In this lecture we will

- learn about simple linear regression and when it is appropriate to carry out this technique.
- learn regression terminology such as residuals.
- learn about the assumptions of simple linear regression
- learn how to make predictions

Prediction and Linear Regression

Now that you have learned how to compute the degree to which two variables are related to one another (correlation coefficient), we can also use these correlations to predict the value of one variable based on the value of another.

This is a very special case of how correlations can be used.

We can only carry out linear regression if the data set is at least moderately correlated i.e. has an absolute correlation coefficient of 0.4 or greater.

Let's consider an example

First Year University Marks	Second Year University Marks
70	66
50	44
80	70
76	54
56	70
38	40
64	62
74	78
54	38
66	74

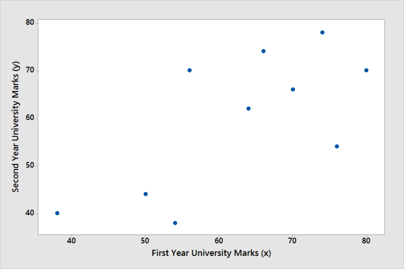
Introducing Regression

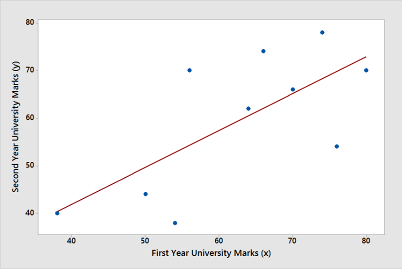
To predict second year university marks from first year university marks we have to create a **regression equation** and then use that to plot a **regression line**.

A regression line reflects our best guess as to what value on the y variable (second year university marks) would be predicted by a score in the x variable (first year university marks). The regression line is drawn so that it minimises the distance between itself and each of the points on the predicted $(\hat{y})$ variable.

Scatter Plots

The left plot is simply a scatterplot of the data and the right plot is a scatterplot of the data with the regression line superimposed.





What does the regression line represent ?

- It’s the regression of the y variable on the x variable. i.e. y (second year university marks) is being predicted from x (first year university marks)
- the regression line is also called the line of best fit - the line fits these data because it minimises the distance between the individual point and the regression line.
- the distance between each individual data point and regression line is the **error in prediction**, this is a direct reflection of the correlation between the two variables.
- if the correlation was perfect all the data points would align themselves along a 45 ° angle.

Terminology

The simplest way to think about predictions, is that you determining the score on one variable (y - **dependent variable**) based on the value of another score (x - **independent variable**).

A **residual** is the vertical distance between an observed value of the dependent variable (y) and the predicted value ($\hat{y}$) on the regression line.

Linear regression model (line of best fit)

The linear regression model (line of best fit) is simply the equation of a straight line.

$$\hat{y} = bx + a,$$

where $\hat{y}$ is the predicted score of y based on a known value of x, b is the slope of the line of best fit, a is the y intercept and x is the score being used as a predictor.

Equations

We will look at these equations when we look at Written Statistics in a few weeks time, so they are in for completeness. You do not need to remember these as they are on the class formula sheet.

The formula for the slope of the regression line (b) and the y intercept, (a) are shown below:

$$b = \frac{S_{xy}}{S_{xx}}, \qquad a = \frac{\sum y - (b \sum x)}{n}$$

Relation to Variance

You will have noticed that S_{xy} and S_{xy} appear in the formula for the slope of the regression line. The formula have been put in for completeness.

You can think of S_{xy}, S_{xx} and S_{yy} as being similar to variance. The formula for S_{xy}, S_{xx} and S_{yy} is very similar to the formula for variance.

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}, \qquad S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Equations cont...

We will look at these equations when we look at Written Statistics in a few weeks time, so they are in for completeness.

Each data point can be be said to have the value

$$y_i = bx_i + a + \varepsilon_i, \qquad \text{where } \varepsilon_i = y_i - \hat{y}_i.$$

here ε_i denotes the residuals.

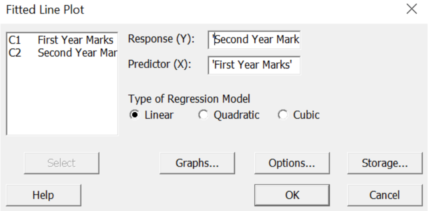
Equation Overload

Try not to worry about all of those equations for just now.
The main one you will need for the project is the regression model i.e.
$\hat{y} = bx + a$.

Linear Regression in Minitab

Let’s now go through the steps of carrying out a regression in Minitab

Stat > Regression > Fitted Line Plot.
Remember your data must be correlated before you can carry out this step.



Fitted Line Plot

C1 First Year Marks
C2 Second Year Mar

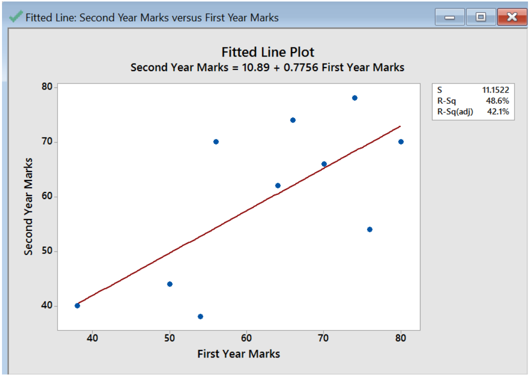
Response (Y): 'Second Year Mark

Predictor (X): 'First Year Marks'

Type of Regression Model

● Linear ○ Quadratic ○ Cubic

Linear Regression in Minitab cont...



We get the following regression line: $\hat{y} = 10.89 + 0.77562x$, where y is second year university mark (%) and x is first year university mark (%).

Interpreting the Regression Line

- Interpreting the Slope
  If your Year 1 mark increases by 1 %, on average the Year 2 mark will increase by 0.77562 %.
- Interpreting the y intercept
  If someone achieves a first year mark of 0 %, i.e. $x = 0$ then their second year mark would be 10.89 %.

Sometimes the interpretation of the y intercept can be meaningless so make sure you check that the interpretation makes sense. In the above example it is unlikely that someone who got a first year mark of 0 % would be allowed to continue on to second year.

How good is this equation ?

Coefficient of determination (R^2): the amount of variance in one variable that is accounted for by the variance in another variable.

R^2 is a proportion and it must take a value between 0 and 1. The higher the number the better the model is. Minitab reports R^2 as a %.

$$R^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}}$$

Interpretation of the Coefficient of Determination

The coefficient of determination is estimated as 48.6 %, meaning that 48.6 % of the variation in the second year marks explained by the linear relationship with the first year marks.

In this class we will only use the R^2, you will learn about the R^2 adjusted in future statistics classes.

Assumptions of the Linear Regression

As with all statistical methods and concepts linear regression has a set of assumptions and they all relate to the residuals.

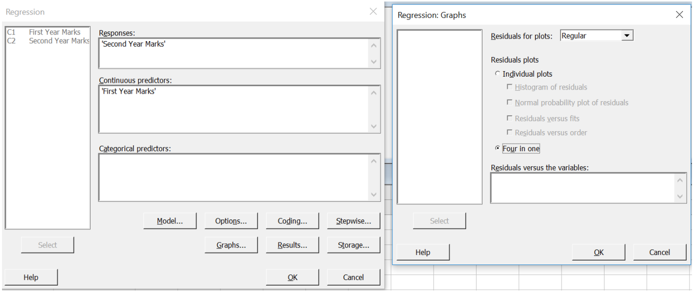
Assumptions of a Simple Linear Regression

1. The residuals should be normally distributed.
2. The residuals should have a mean of zero.
3. The residuals should have constant variance (another word for constant variance is homoskedastic).
4. The residuals are uncorrelated with each other.

Testing the Assumptions in Minitab

We test the assumptions of a linear regression by using a **4 in 1 diagnostic plot**.

Stat > Regression > Regression > Fit Regression Model. The press on Graph and select 4 in 1, as shown in the image on the right.



Regression

C1 First Year Marks
C2 Second Year Marks

Responses: 'Second Year Marks'

Continuous predictors: 'First Year Marks'

Categorical predictors:

Regression: Graphs

Residuals for plots: Regular

Residuals plots

○ Individual plots

□ Histogram of residuals

□ Normal probability plot of residuals

□ Residuals versus fits

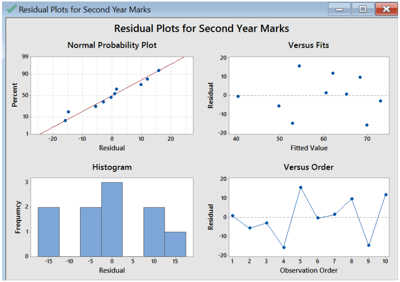
□ Residuals versus order

● Four in one

Residuals versus the variables:

Testing the Assumptions in Minitab cont...

This is what your 4 in 1 diagnostic plot may look like.

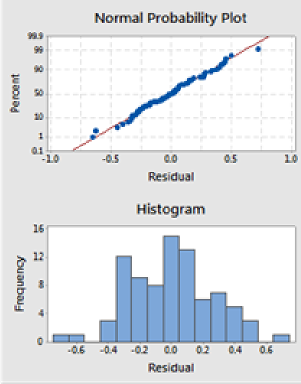


The next slides will discuss what you are looking for in each plot and what assumption is being tested.

Checking the Assumptions

What assumption is being checked ?
The residuals are normally distributed.
What are we looking for ?

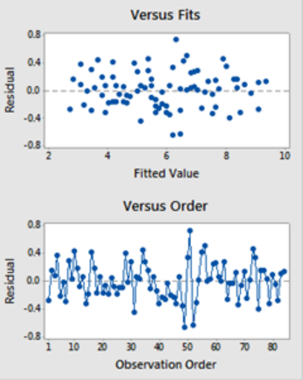
1. In the normal probability plot the majority of the blue points should lie along the red line.
   Chubby pencil test - if you put a chubby pencil over the red line do most/all of the blue dots disappear ? If yes then we can assume the residuals follow a normal distribution.
2. Histogram should be roughly symmetric, remember this is subjective so the Normal Probability plot is more informative.



Checking the Assumptions cont...

What assumption is being checked ?
The residuals are uncorrelated.
What are we looking for ?

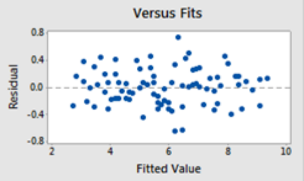
1. In both plots there should be no obvious trend in the pattern of the residuals and we want a random scatter of points.



Checking the Assumptions cont...

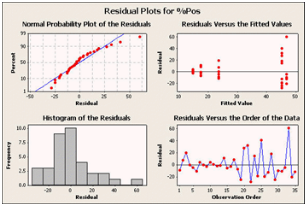
What assumption is being checked ?
The residuals have constant mean zero.
The residuals have constant variance.
What are we looking for ?

1. There will always be a horizontal line at 0.0 for this mean to constant we want a random scatter on either side of the dotted line.



What you don’t want to happen.

- **Normal Probability plot**
  Does not follow a normal distribution as the points do not lie on the line
- **Histogram**
  The histogram is right tailed skewed.
- **Residual Versus Fitted Value**
  There is a trend - the residuals are fanning out.
- **Residual Versus Order**
  1st half - correlated
  2nd half - uncorrelated (variation is greater).



What to do if the assumptions are not met

If you find the assumptions are not met, you just need to comment on this.

This means that although the data was correlated carrying out a simple linear regression is not appropriate.

Predictions

If you wanted to predict what second year university mark a student would get if they got 55 % in first year, then

$$\begin{aligned}\hat{y} &= 10.89 + 0.77562x \\ &= 10.89 + (0.77562 \times 55) \\ &= 10.89 + 42.6591 \\ &= 53.5491\end{aligned}$$

Therefore we predict that if a student had a first year mark of 55 % then we predict a second year mark of 53.5 %.

You cannot make predictions outside the range of your data.