Scatter Plots, Regression Lines and Residuals

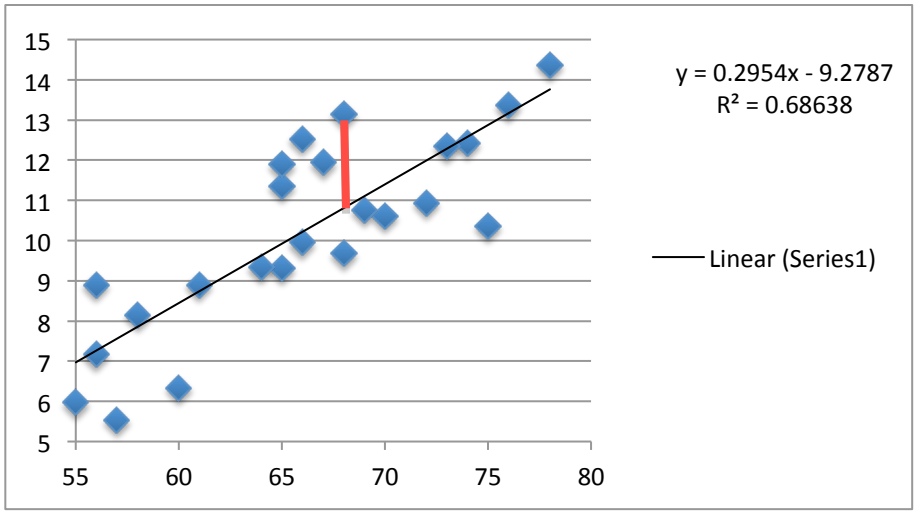
-A **scatter plot** is a graphical representation of the relationship between two quantitative variables on the same sample. It is formed by plotting the pairs of measurements from all of the units in the sample on a corresponding pair of axes representing the two measurement scales.

-The **regression line** associated to a scatter plot is the line that lies overall closest to the points in the scatter plot (technically, it minimizes the squares of vertical distances of points in the scatter plot to the line). It can also be thought of as giving the best linear fit for the relationship between the two variables.

-The **residuals** are the set of vertical distances from the points in the scatter plot to the regression line, and represent the differences between the values of the response variable predicted by the regression line and the measured values of the response at particular values of the predictor.

Example: The plot below represents the relationship between the variable height in inches (the horizontal axis) and the variable weight in stones (vertical axis) for a set of British students. Each blue diamond represents the height and weight measurements from one student. The black line is the regression line. The equation of this regression line is given to the left, y=0.2954x -9.2787, where y=weight and x=height. Also given in this diagram is the **coefficient of determination**, R^2, which expresses the strength and direction of the relationship between the two variables, and is the square of the correlation coefficient, R. The farther R^2 is from zero (either positive or negative), the stronger the relationship.

One residual is shown below in red, betwen the point with coordinates height=68 and weight=13.2 and the regression line, which predicts a weight of 10.8 stone for a height of 67cm. The value of the residual is the diference between the weight given by the datapoint and the weight given by the residual line for a given height. In this case the residual is 13.2- 10.8 = 2.4.

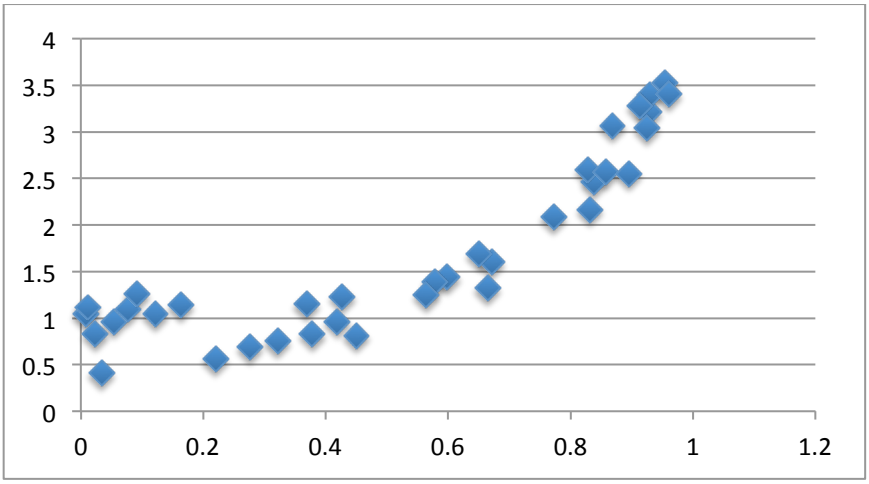


Examining a Regression Line Fit

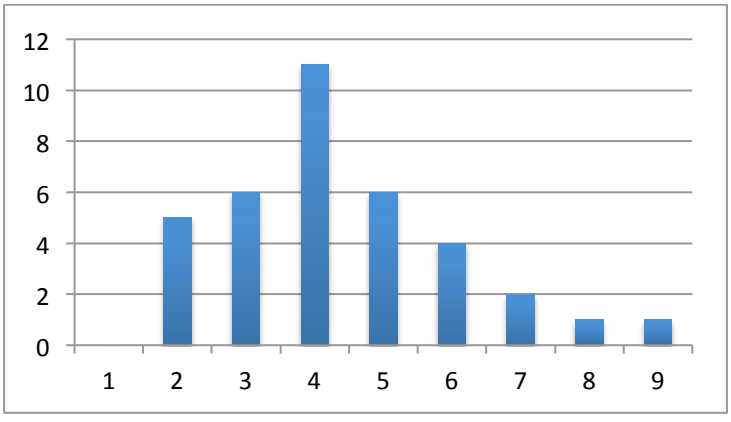
There are three checks you should make about a regression line fit.

1. **General quality of fit:** visually examine the scatter plot to see if a linear fit is appropriate or if the data should be modelled by a different curve. In the scatterplot above, the data is well approximated by a line.

You should be on the look out for plots such as the one below, which seem to curve rather than following a straight line. In such a situation, you may need to consider a more complicated relationship between your variables for your analysis. Recent research in statistics suggests that it is better to approximate a nonlinear relationship between variables by splines than polynomials.

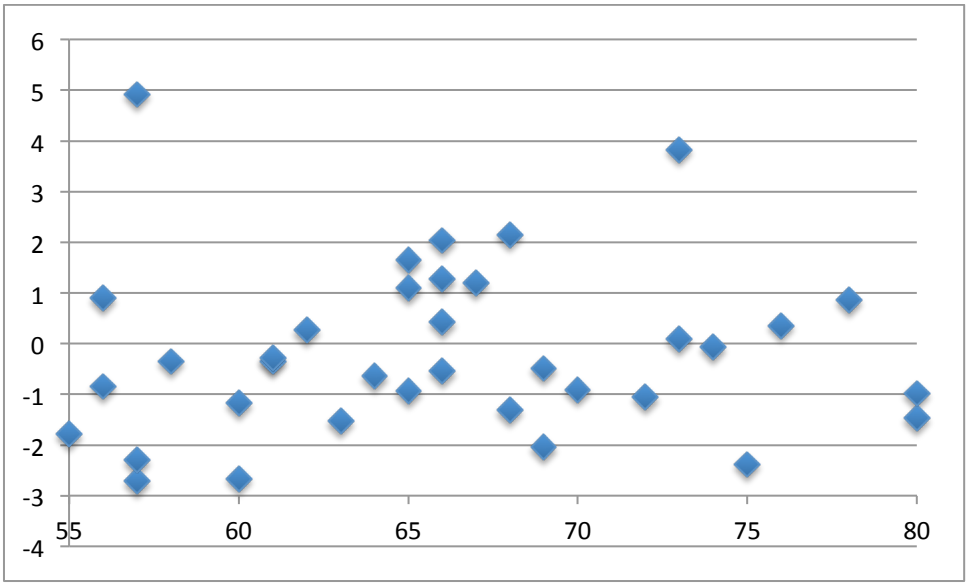


2. **Normality of the residuals**: Calculate the residuals and plot a histogram (or bar chart as shown blow) to visually evaluate the normality. The residuals below are slightly skewed, but relatively normal, so the assumption of normality of residuals is reasonable. If you have a fairly small dataset, it may be difficult to evaluate the normality of your residuals through a histogram. In that case, you may wish instead to use a Q-Q plot to examine normality. If you have a large dataset, due to the normalising effect of the Central Limit Theorem, it may be okay to run a test assuming normal residuals even if your residuals are not normally distributed.

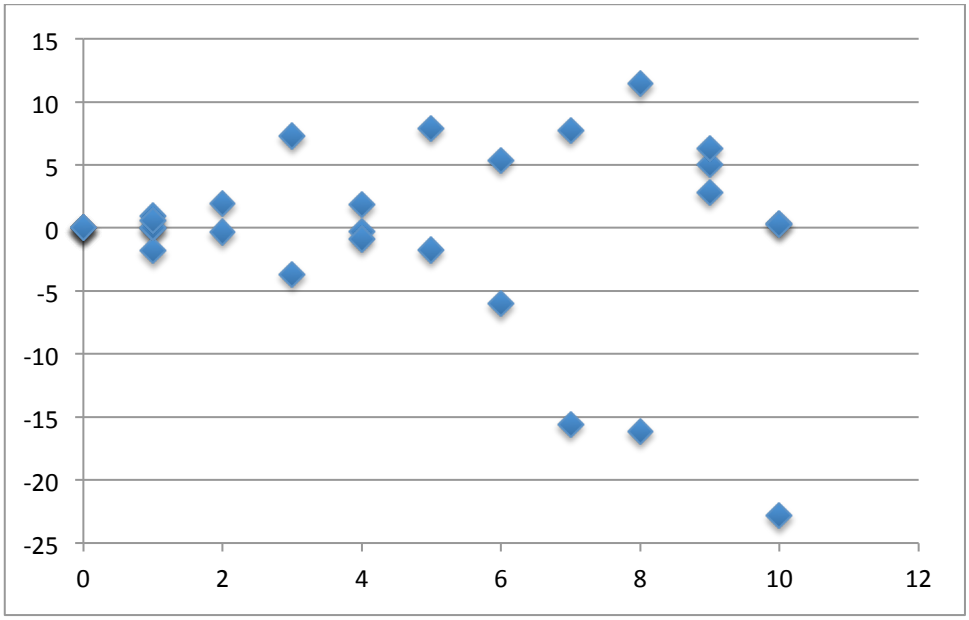


3. **Homoscedasticity of the residuals**: for each experimental unit, plot the point with coordinates given by the value of the predictor variable and the residual for that point. Examine visually that the plot looks roughly flat.

Below is the plot for the scatter of heights and weights. This looks pretty level, so the assumption of homoscedasticity is okay.



Compare this with the following residual plot, in which the residuals spread out more as the predictor variable increases. The residuals in this second plot do not satisfy the assumption of homoscedasticity. Such residuals are called heteroscedastic.



In general, heteroscedasticity of residuals is considered a serious problem as they can cause problems with analysis and may result in lower p-values from the analysis than are really justified; that is, you may obtain incorrectly significant results if your residuals are not homoscedastic. A general rule of thumb is that if the variance of

the residuals for the top 10% of predictor values is more than four times the variance of the residuals for the bottom 10% of the predictor values, you need to resolve this issue before continuing with your analysis.

There are three general ways to deal with heteroscedasticity of residuals.

1. **Transform your data**: Often problems with homoscedasticity and normality occur together and can both be resolved by an appropriate transformation.

2. **Use weighted least squares to fit your regression line**: This has been the conventional way to deal with heteroscedasticity in the past. This method can be difficult to apply, so it is advisable to see a consultant for advice.

3. **Use bootstrap methods**: Bootstrap methods are probably the best way to deal with heteroscedasticity if transformation does not work, but again, can be difficult to apply, so it is advisable to see a consultant. In addition, as they are newer techniques, they may be less recognised in your discipline.

As always, it is a good idea to look at how analysis has been done on the variables you are using, as well as how the issue of heteroscedasticity has been dealt with previously in your discipline before deciding on the method to use in your analysis.