Statistical literacy guide
A basic outline of regression analysis

Last updated: March 2009
Author: David Knott & Paul Bolton

Social scientists are often interested in analysing whether a relationship exists between two variables in a population. For instance, is greater corruption control associated with higher GDP per capita? Does increased per capita health expenditure lead to lower rates of infant mortality? Statistics can give us information about the strength of **association**, this can sometimes help in deciding if there is a **causal** relationship, but they are not sufficient to establish this on their own.

Relationships are often expressed in terms of a **dependent** or response variable (Y) and one or more **independent** or describing variables (X_i). The dependent variable is the condition against which certain effects are measured, while independent variables are those that are being tested to see what association, if any, they have with the dependent variable.

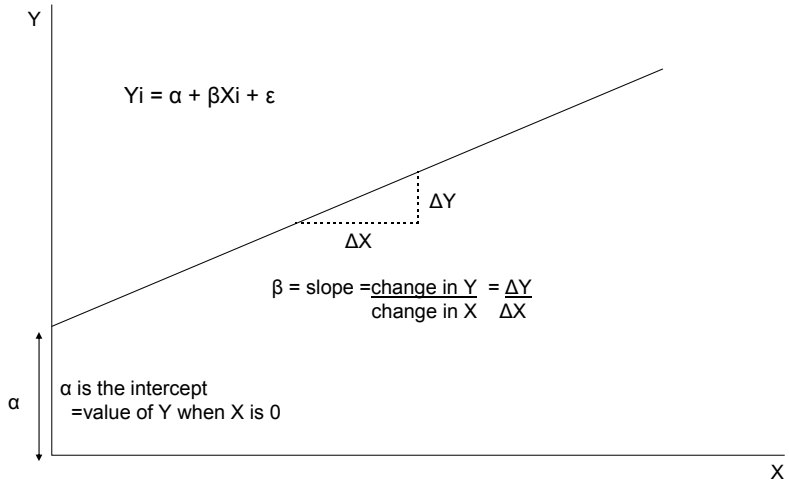
A scatter plot of Y against X can give a very general picture of the relationship, but this is rarely more than a starting point. Statisticians can establish the direction, size and significance of a potential association between an independent and dependent variable using a **simple linear regression model**.

- **Simple** because one explanatory variable is being tested, rather than several
- **Linear** because it is assessing whether there is a straight-line association between X and Y
- **Model** because the process is a simplified yet useful abstraction of the complex processes that determine values of Y,

The typical notation for a simple regression model is in the form of an equation of a straight line:

$$Y_i = \alpha + \beta X_i + \varepsilon$$

Where Y_i is the dependent variable, X_i the independent variable, and α is the intercept (i.e. the value of Y when β equals zero). β is the regression coefficient – the slope of the line that shows the increase (or decrease) in Y given a one-unit increase in X_i. These parameters are estimated in order to reduce ε^2 and produce a line of best fit. The elements of this equation are illustrated below:



A number of statistical tests can be conducted to measure the strength of the association. Normally the most important model parameter to test is *β* the regression coefficient. As with Confidence intervals and statistical significance the starting point is hypothesis testing.

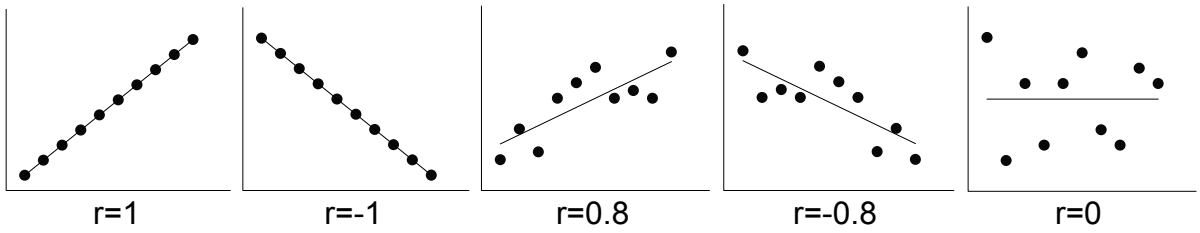
Hypothesis:

H_0 : $\beta = 0$ There is no (linear) association between X & Y in the population

H_a : $\beta \neq 0$ There is some (linear) association between X & Y in the popn

When a researcher rejects the null hypothesis, what is being said is that probability of observing a test statistic as large as that observed in the standard test statistic - with the null hypothesis being true - is so small enough that the researcher feels confident enough to reject it. This probability (often known as a significance level) is often 0.05, or 0.1, but can be as low as 0.01 (in medical trials).

The Pearson product moment **correlation coefficient** (or simply the correlation coefficient) is not a model parameter, but a measure of the strength of the association between the two variables. The correlation coefficient is denoted *r* and can take values from -1 to 1. A positive figure indicates a positive correlation –an upward sloping line of best fit, and *vice versa*. If r=0 then there is a complete lack of linear correlation. In practice such extreme results are unlikely and the general rule is that values closer to +1 or -1 indicate a stronger association. Some examples are illustrated below.



R^2, sometimes known as the coefficient of determination, measures the proportion of the variance in Y that is explained by X. It has no sign, so values closer to 1 indicate a closer association –that X is better at predicting Y. R^2 is sometimes given as the sole or most important measure of the association between X and Y and hence the usefulness of the model. However, a model that has a high R^2 is not likely to be useful if we can not reject the null hypothesis $\beta = 0$. The interpretation of a particular value of R^2 is not purely statistical. A high value does not necessarily means that X causes Y or than it is a meaningful

explanation of Y. It depends on the nature of the variables being looked at. Associations in the social sciences tend to have smaller R^2 values than those in the physical sciences as they are dealing with human factors that often involve more unexplained variation. The R^2 value is not a measure of how well the model fits and a useful model can have a low value.

Association and causation

In his classic essay on causation Sir Austin Bradford Hill set out his views of the most important factors to consider when deciding whether an observed statistical association is due to causation.[1] These are given in descending order of importance:

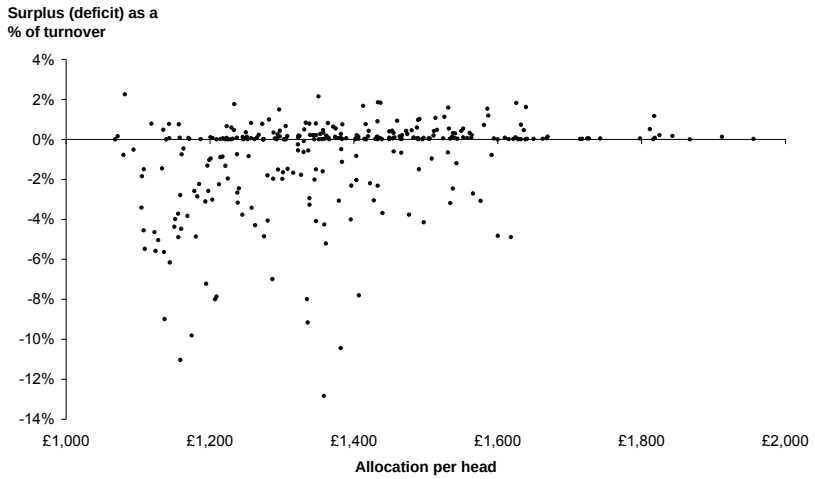
1) *Strength* - What increase in cases of the potential effect or outcome is observed when the potential cause is present? Strength here refers to differences in the instances of the effect or outcome, not the statistical strength of any association which has to be 'significant' and not down to chance before looking at a hypothesis of causation.
2) *Consistency* -Has the finding been repeatedly observed, by different people, at different times and under different circumstances?
3) *Specificity* –How specific is the potential effect? Is it limited to particular groups? Is the potential cause associated with other outcomes? A high degree of specificity can lend great support for a causal hypothesis, but such clear, simple and distinct one-to-one relationships are rare.
4) *Temporality* -In what order did the event happen? An effect needs to come after a cause.
5) *'Biological gradient'* –Is the effect stronger where the potential cause is stronger (more intense, longer duration of exposure etc.), a so-called dose-response curve?
6) *Plausibility* -Is there a plausible theory behind the hypothesis of causation?
7) *Coherence* -Does the hypothesis make sense given current knowledge and related observations?
8) *Experiment* -Is there any experimental evidence specifically connected to the hypothesis?
9) *Analogy* –Are there any similar causal relationships?

Example: Primary Care Trust deficits and health allocation per head

You have obtained data for all Primary Care Trusts. Data include outturn as a proportion of PCT turnover and health allocation per head of population. How can you tell whether there is an association between the variables?

The scatter plot below suggests a possible positive relationship which makes some intuitive sense, but the points do not appear to form anything like a straight line, so we can expect that only a small proportion of the variation in Y is explained by X. We can not draw any firm conclusions without calculating and testing the model parameters.

[1] Austin Bradford Hill, *The Environment and Disease: Association or Causation?*, Proceedings of the Royal Society of Medicine, 58 (1965), 295-300.
Reproduced at: http://www.edwardtufte.com/tufte/hill



Regression model:

$$Y_{ot} = \alpha + \beta X_{all} + \varepsilon$$

Where Y_{ot} is the PCT outturn expressed as surplus/deficit as a % of turnover, X_{all} the allocation per head and α is the intercept (i.e. the outturn level if funding per head is zero).

Null hypothesis

$H_0 : \beta = 0$ There is no (linear) association between PCT outturn and allocation per head in the population

Alternative hypothesis

$H_a : \beta \neq 0$ There is some (linear) association between PCT outturn and allocation per head in the population

The model parameters and other regression statistics can be calculated in Excel (Tools; Add-ins; tick Analysis Toolpak; then back to Tools; Data Analysis; Regression)

Regression output format should be similar to this.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.30713875
R Square	0.094334211
Adjusted R Square	0.091325355
Standard Error	0.022163667
Observations	303

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.015401074	0.015401074	31.35218091	4.85091E-08
Residual	301	0.147859674	0.000491228		
Total	302	0.163260748			

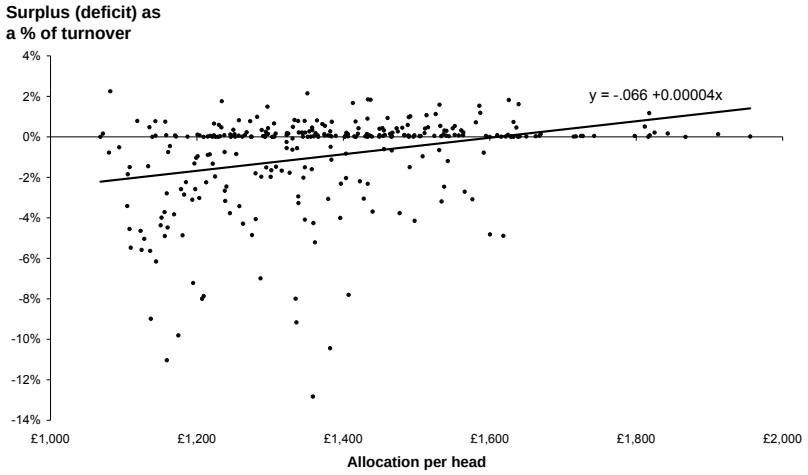
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.065670268	0.010129681	-6.482955226	3.67695E-10	-0.085604247	-0.045736289
X Variable 1	4.07875E-05	7.2844E-06	5.599301823	4.85091E-08	2.64527E-05	5.51223E-05

Interpreting the output

The coefficients column enables us to state the regression model. In this case:

$$Y_{ot} = -0.066 + 0.00004X_{all} + \varepsilon$$

Each £1 increase in health allocation is estimated to result in a 0.00004 increase in PCT outturn as a proportion (%) of turnover. Alternatively a £100 increase in allocation per head is estimated to result in an increase in turnover of 0.4 percentage points. This line is illustrated in the next chart.



The test of statistical significance is to calculate a t-statistic (given in the regression output above). The probability (P-value) of observing a t-test statistic as high as 5.6 gives the answer to the hypothesis test:

$H_0 : \beta = 0$ is <0.001 (P=0.0000005)

The probability that there is no linear association is extremely low so **we can therefore reject the null hypothesis.** There is therefore some linear association between PCT outturn as a % of turnover and health allocations per head. The output shows a positive linear association between outturn and allocation per head among the population (what we expected before the analysis). As the P-value is so low this relationship is significant at the 5% (or even 0.1%) level of significance. The 95% **confidence interva**l for β is also given in the regression output; 0.000026 to 0.000055. Alternatively the confidence interval of an increase of £100 in allocation per head is 0.26 to 0.55 percentage points.

In this case R^2 equals 0.09 – suggesting that 9% of the variation in PCT outturns as a proportion of turnover are explained by changes in allocation per head. This is a low degree of explanation and confirms the visual interpretation of a wide spread of results around the regression line. There is a general positive association, but allocations per head explain little of the variance in levels of deficit. The model therefore does not help to explain much of what is going on here and is poor at predicting levels of deficit from allocations. This might imply that other explanatory variables could be usefully added to the model. Where more variables are added, this is called “multiple regression”.

Other statistical literacy guides in this series:

- What is a billion? and other units
- How to understand and calculate percentages
- Index numbers
- Rounding and significant places
- Measures of average and spread
- How to read charts
- How to spot spin and inappropriate use of statistics
- A basic outline of samples and sampling
- Confidence intervals and statistical significance
- A basic outline of regression analysis
- Uncertainty and risk
- How to adjust for inflation