# Week 8

Refer to the course map

This week we conclude our exploration of inferential tools with another very common application: regression analysis, and study of correlation. This is a rich source of misunderstanding, so we want to be clear and focused. The most obvious source of misunderstanding is possible confusion between "correlation" and "causation": the fact that two things seem to come in pairs, does not mean at all that one is the, or even a, *cause* of the other. The classic example (even if a very biased one) is the "correlation" between smoking and lung cancer: statistical evidence is unequivocal that the two are correlated, but statistics alone does not *prove* that smoking *causes* cancer. Hypothetically (and this has been an argument used by tobacco companies for a long time), there could be some "hidden" common cause that pushes people to smoke, and to get lung cancer (even it would stretch people's imagination, as far as the statistical evidence goes, one could even argue that it is a strong propensity for lung cancer that causes people to smoke, even if it is hard to figure how this could be). The chemical investigation identifying well-known carcinogen in tobacco smoke is the clincher in the argument that smoking is indeed a cause of lung cancer. Without this additional hard evidence, the argument against tobacco would be much weaker. In other circumstances, where there is no similar supporting, and independent evidence, proof of *correlation* is absolutely no evidence of *causation*. Whenever someone argues differently, they are relying on additional (explicit or hidden) arguments.

A less obvious problem is under what conditions a model like the one discussed here will hold, and what it would mean (and *not* mean) when they don't hold. We don't have much time to dwell on this issue, but we can have a first good look at the various issues.

## Objectives for this week

* Recognize various levels of theoretical justification for a linear regression model
* Be able to calculate regression coefficients, and related quantities using a spreadsheet

## Reading Material

We look at the theoretical underpinnings of *linear* regression , and the study of correlations. (here is the PDF version) We also take a brief visit to some *nonlinear* cases, which, however,

* reduce, in a sense, to linear (if they don't, they are not easily studied)
* do not have as solid a theory behind them, but are rather "ad hoc" models

The discussion above should help you distinguish between one class of regression models (when strong assumptions are made on the pair of variables under scrutiny), and the more generic least mean square approach. In the latter case, linear models are a choice of convenience and simplicity, but we don't have solid arguments to assert their validity for prediction purposes (which is actually the whole point of this topic).

This topic has many technical aspects, and in the unit linked above, we are really more concerned about its foundations rather than the details of its implementation. While we do not have time to go over the latter in great detail, as in lots of exercises and applications, the following links will introduce you to this side of the topic (all of the calculations are, as usual, hard-coded in most spreadsheets, and definitely in all higher level statistical packages - it's the scope of this theory that might not be as well covered). You should check out the following material:

* Correlation theory (its strong foundation is when the variables are jointly Gaussian - in all other cases it is a measure of *linear* dependence, except that is not as significant, when normality is not available). Please, note that "Pearson's *r*" is the correlation coefficient we defined in the file on linear regression, as the *empirical* version of the theoretical correlation. A brief discussion of the sampling distribution for *r* is here. This can be used to build (approximate) confidence intervals
* The corresponding linear regression theory is linked here, with the explicit formulas to evaluate the errors involved (again, these formulas have a direct meaning in the Gaussian case - in general, they obviously measure something, but it is not as a specifically pertinent probabilistic object as in the Gaussian case)

## Assignments

We apply regression methods to our simulated regression data (from week 1). Compare the normal data with the non normal data, and discuss how the two compare, in particular, concerning the resulting sum of residuals .

**Please, turn in your work by Monday, Week 9**

### Quizzes

You can get credit for participation by taking the quizzes 66 to 75 on the [WAMAP](http://www.wamap.org) site: this group is about regression and correlation between variables.