Question Week 1
Johannes Hengelbrock (2206164)
06.09.2012

During the last lecture, we talked about ways to decrease the standard error of a regression coefficient (e.g. by increasing the sample size). If I remember it correctly, you said that the standard error of a point estimate can be decreased by increasing the overall model fit (R^2), no matter how.

I have a hart time understanding why this is the case.

Lets assume we want to research the effect of X on Y (lets call the effect „β_k“) by using a simple regression. Now, suppose we found a third variable, Z, that is completely uncorrelated with X but explains a lot of variation in Y (=increases R^2). If I understood you correctly, including Z in the regression would decrease the standard error of β_k, although the point estimate is not affected (since X and Z are uncorrelated). I think you are right, because the formula for the standard error is the following:

standard error of β_k $$= \sqrt{\frac{1-R^2_{YH}}{(1-R^2_{X_kG_k})*(N-K-1)}} * \frac{s_y}{s_{X_k}}$$

with:

H = the set of all the X (independent) variables (in this case X and Z).
G_k = the set of all the X variables except X_k (in this case Z).

(I got the forumla from this page: http://www.nd.edu/~rwilliam/stats1/x91.pdf , why the notation is a bit ugly)

Now, because X and Z are completely uncorrelated, the term R^2_{XkGk} becomes 0. This term indicates the size of multicolliarity and it makes sense that the bigger this term, the larger the standard errors are. R^2_{YH} stands for how much variance in Y can be explained by the model, so by <u>all</u> variables. Now, if we would estimate the simple regression of X on Y, the standard error of β_k would be larger than if we would include Z (because the multicolinearity stays at 0 and the overall R^2 increases).

My question is: why should that be? Why does including Z decrease our uncertainty about β_k, even though X and Z are completely uncorrelated (do not provide any information about each other)?

Also if I think about the consequences of this: suppose we have a medical trial in which a drug is randomly assigned to a sample and we want to test how the drug performs compared to taking no drug (or a placebo; say Y is whether a patient dies or not). Because assignment is random, taking the drug is by design uncorrelated with everything. Now, assuming the discussed above is true, wouldn't that mean that we could increase our certainty about the effect of the drug by including predictors of Y such as age, gender, some genetic stuff and so on? Because they would not be correlated with the assignment to the drug but increase the overall model fit? If this is true, is this actually done in medical research?