METHODS OF QUANTITATIVE DATA ANALYIS
MSR Course, 2011-2012

Harry B.G. Ganzeboom

Lecture 2.1: Simple and Multiple Regression

April 12-16 2012

Agenda

- Causal model diagrams, confounding, indirect and interaction effects
- Assignment 1: Causal analysis with table elaboration.
- Simple and Multiple Regression (DJT, ch 5-6)

Linear models

- Linear models (Y = a + b*X) are the expression of the most common scientific argument: if X causes Y, we expect that more X increases Y.
- Even if such statements can be interpreted ordinally, we routinely use an interval model to test it.
- The linearity of the model is in practice quite flexible and can accommodate:
  - Ordinal X and ordinal Y,
  - Discrete (nominal) X,
  - Non-linear relationships of various kinds, including non-monotonic (e.g. U-shaped relationships).
  - Multiple X variables.
  - Extension to nominal Y are straightforward, but not trivial.
- Almost all models for whatever are a variation or extension of the simple linear (and additive model): GLIMs.

Regression and OLS

- The linear model can also accommodate that X is not the only cause of Y, and is then phrased as: Y = a + b.X + residu.
- A commonly used alternative term for *residu* is *error*. This is fine, a long as you understand that this does NOT mean that the model is *wrong*.
- As the relationship between X and Y is probabilistic, we need to find values for a and b, that produce a *best fitting* line.
- The most often used criterion to estimate a and b is that of minimizing the sum (or average) of squared residuals (SS-res of MS-res): OLS.

The OLS solution

- The coefficients of the simple regression equation are obtained as (DJT, 91):
  - Slope: b = cov(X,Y) / var(X).
  - Intercept: a = mean(Y)-b*mean(X).
- Forget about the computational formulae.

Why least squares (DJT, 91)?

- It turns out that the least squares criterion is algebraically more convenient than (e.g. absolute deviations). In particular it leads to closed form solutions (no iterations) and additive decomposition of variation.
- It can also be show that OLS is – asymptotically -- equivalent to other reasonable solutions, in particular Maximum Likelihood – which generally requires iterative procedures.

Consequences of least squares

- Least squares estimates give greater weight to outliers – and they can be sensitive (have greater variability) when you are sampling from distributions on which outliers can occur.
- An obvious remedy is omit outliers from you sample, which is sometimes known as *robust* regression – but not often used.
- We will learn more about outliers and influential cases in the future.

Why “regression”?

- Why are these models called “regression” models?

→ These models were first introduced by bio-statisticians who studied the relationship between generations of the same characteristic (e.g. in height).

- The coefficient is such problems measure to what extent the second generation is not equal to the first generation, but has “regressed towards the mean”.
- So it is not a very helpful name!
- Better would be: “linear model”, “additive model” for the form of the model, and “OLS model” for how it is estimated.

Explained variance (DJT, 91)

- A best fitting (regression) line is found by minimizing the SS-Error (sum-of-squared errors): SUM(y-^y)^2.
- SS-total = SS-model + SS-error.
- SS-total = SUM(y-mean(y))^2.
- SS-model = SUM(^y-mean(y))^2.
- Explained variance is a proportion: SS-model / SS-total.

Testing the significance of the equation

- SS are also the ingredients of the overall F test of the equation:
  - F = MS-model / MS-error.
- In which MS are obtained by dividing SS by the associated degrees of freedom:
  - N total number of cases
  - K number of estimated effects in model
  - N-k-1 residual degrees of freedom
- The F-test is usually not so interesting, but gains importance when there multiple X-variables and you are comparing models.
- You should be able to read the ANOVA tables in regress.

Correlation

- (Multiple) correlation R denotes the fit around the regression line. It is obtained as the sqrt(R2).
- (Simple) correlations can also be defined as standardized covariation: r = cov(x,y) / sqrt(var(x)*var(y)).
- Simple correlations have a +/- sign and range between –1.00 and +1.00.

Standardized regression

- If we express X and Y in standardized terms (z-scores), we obtain the regression equation in standardized form:
  - Z(Y) = 0 + beta*z(X)
- SPSS prints these beta's rountinely (in regress), Stata does not.
- Standardize regression coefficients are in the same metric as correlation coefficients; in simple regression they are identical.
- Beta = B*(sd(x)/sd(y)) and B = beta*(sd(y)/sd(x))

Multiple regression

- Multiple regression refers to multiple X-variables.
- It is NOT the same as multi*variate* regression (which would be multiple Y-variables).
- Y = b0 + B1*X1 + B2*X2 + … + residu.
- With two X-variables the model implies a plain in three-dimensional space (DJT, 105), but with more X-variables the geometry becomes (even) less helpful.

Partial interpretation

- B-coefficients in multiple regression have a partial interpretation: Bk informs about how Xk changes Y, controlling all other X-variables in the model.
- Multiple regression is a restatement of what we see in table elaboration.
- However, it also works when X-variables are not discrete (but continuous), nominal, and it also works if we have many X-variables.

→ with regression models available, there is no need for tabular analysis.

How does multiple regression control?

- Y = B0 + B1*X1 + B2*X2
- Regress: Y = B1*X1, take RES1
- Regress: Y = B2*X2, takes RES2
- B1 can be obtained in Y = B1*RES1
- B2 can be obtained as Y = B2*RES2

→ Multiple regression coefficients are the effects of Xk of residualized Y-var.

The F-test

- In multiple regression, the F-test in Anova become a bit more interesting, as it can be transformed in an F-change:
  - F-change = ((SS-model1 – SS-model2)/df) / MS-error
- DJT, 124 present a formula using R2. This is the same. You can ask SPSS to calculate the F-values.
- This tests the statistical significance of the additional explained variance, which may be different from the significance B-coefficinest of the added variables.

Dummy variables

- Multiple regression can be applied to model effects of a nominal X, expressed in indicator (0/1) dummy variables.
- Dummy variables look like different variables, but they are not.
- The full set of dummy variables is linearly dependent, we have to omit one to obtain a *reference* category.
- R2 in (simple) dummy is also know as the correlation ratio (DJT, p. 99).

The reference category

- Always, always inform the reader about the omitted category (unlike DJT, 127).
- Never choose a very small category as the reference.
- It is better to be in control of the choice, but even better to check by entering “all” dummies. In particular with missing values in your data, or with multiple dummy sets, this can give you results to think twice.

High or low R2?

- Researchers often seem to interested in obtaining high explained variance.
- This is the wrong attitude. You should ask yourself whether a model adequately represents the (causal) relationships that you are studying.
- In this sense it is incorrect to call R2 a fit-statistic.
- Adequate question to ask are:
  - Is a linear specification correct? This can be tested.
  - What could be omitted (confounding) variables?
- Note that omitting variables that (strongly) determine Y, but do not determine X, does not invalidate the model!!!
- If you omit such predictor variables, this will only effect statistical power.
- However, if you omit confounders, this invalidates the model.

Test of linearity

- A simple test of linearity can be obtained by categorizing a continuous variable into discrete categories and compare the continuous variable with the dummy variable model.
- It is convenient when you can easily shift between a categorical and a continuous representation of a model. In Stata this can be done with the I.expander, in SPSS in UniAnova by choosing between *by* and *with*.

Not yet discussed

- Factors affecting the size of correlation and regression coefficients (DJT, 94): outliers, leverage points, truncation, regression toward the mean, aggregation.
- Multi-collinearity (DJT, 108).
- SE of the Estimate.
- Regression models with (discrete) interactions (DJT 124).