Class notes 7

Data related issues

Missing data

When working with social science data, one typically run into the problem of missing data observations. For instance, in survey data we often see incomplete answers (items for which the respondent forgot or chose not to answer). One distinguish between MCAR (missing completely at random), MAR (missing at random) and MNAR (missing not at random). Missing data due to MCAR will not cause the statistical analysis to be biased. This is not true for MNAR, which may cause bias when the data is analyzed.

Two main strategies for how to handle missing data:

1. *Listwise deletion*: Works by simply eliminating the entire row in the data table if one or more values are missing from the row. To make this more clear, consider a data table consisting of the variables X_1, X_2, X_3, X_4. Assume that the score on X_2 is missing at row 10. Then by listwise deletion, the 10^{th} row is deleted from the data table. Listwise deletion is the most common strategy, and most software packages are set to handle missing data that way.

2. *Imputation*: Works by estimating the missing data score. The most obvious method of imputation is to substitute the missing score with the sample mean of the variable in question. Methods that are more refined are *regression*, the *EM algorithm* (Expectation-Maximization) and *multiple imputation*.

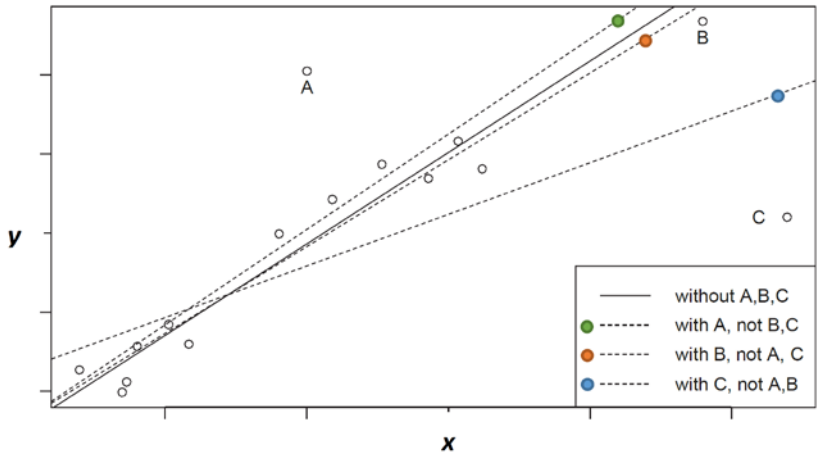
Outliers

An outlier is an observation that is numerically distant from the rest of the data. How to handle outliers in the data is a difficult matter. What is the effect of outliers?

Consider the simple regression model

$$Y = B_0 + B_1 X + \varepsilon$$

Let $\{(y_i, x_i)\}_{i=1}^{15}$ be a set of pairwise realizations of Y and X. Now, consider a plot of the data with the outlying cases indicated by A, B and C (taking from Yuan and Zhong, 2013). The plot show four different regression lines estimated using the OLS procedure. The effect of the outliers is clearly visible.



The usual way to detect univariate outliers is to evaluate the z-scores of the variable in question. The multivariate version of the z-score is the Mahalanobis distance measure.

If it is possible to identify outliers in the data, different strategies may be implemented to handle the problem. One needs to make a careful assessment about the likely cause of the outlying cases. A common approach is to eliminate the outlying observations in the data. Such an approach is controversial since the outlying data points may represent relevant information about the phenomenon under study. If, however, there are reasons to believe that some fault occurred in the process of collecting and registering the data, then (obviously) the appropriate action is to eliminate the faulty observations.

Other common strategies are:

1. *Trimming the data*: a popular strategy is simply to delete the end points of the data. For instance, when the researcher choose to remove the 5 % smallest and largest values of some variable.

2. *Transforming the data*: in some cases, transforming the data will help alleviate the problem. For instance, logarithmic transformation has the effect of "compressing" the scale of a variable. However, our primary concern is to specify a correct model. Thus, transforming the variables in the model for the sake of handling the outlier problem may be bad idea.

3. *Alternative estimation methods*: recall the OLS loss function

$$\min_{b_0,\ldots,b_k} L = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

The OLS estimator is sensitive to outliers since it works on the squared residuals. There are alternative estimation procedures, which may perform better in the presence of outliers. For instance, consider

$$\min_{b_0,\ldots,b_k} L = \sum_{i=1}^{n} |\hat{\varepsilon}_i|$$

This estimator is known as *Least absolute deviation* (LAD).

4. *Ignore the problem*: run the analysis with and without the outlying information. If the conclusions are unaltered from including the outliers, one should not be too concerned about the issue.

Another consequence of outliers, which is often overlooked in textbooks, is that many statistical procedures rely on numerical optimization algorithms. The performance of such algorithms may be sensitive to the presence of extreme observations.

Assumptions underlying the linear regression model

We may ask ourselves what can go wrong when performing linear regression analysis. Recall the assumptions underlying the linear regression model. It is worth discussing these assumptions in some detail.

Assumption: the regression model is linear in and correctly specified

A common way to evaluate the assumption of linearity is to plot the dependent variable against the predictors. If the form of the relationship between the dependent variable and any of the predictors appears non-linear, some suitable transformation may be necessary to resolve the issue (see previous lecture notes regarding functional form).

Assumption: the mean value of ε (given the Xs) is zero.

The assumption is essential for the OLS estimator to preserve important statistical properties such as *consistency* and *unbiasedness*. Recall that consistency is the property that the estimator converge to the true population value as n tend to infinity; and unbiasedness is the property that the estimator on average produce estimates that are equal to the true population values.

The assumption is typically violated when the model is incorrectly specified, which in turn cause the error in the model to be correlated with any (or all) of the predictors. The most obvious form of misspecification is to adopt a wrong functional form when formulating the model. Another misspecification is represented by *omitted variables*. If we, by mistake omit a relevant predictor, and if this predictor is related to the remaining predictors in the model, the estimated parameters will be inconsistent and biased. Consequently, any inference about the relationships in the model will be invalid. A third form of misspecification is so-called *measurement error*. It will have an adverse effect on the quality of the estimated parameters if one or several predictors in the model are measured with error.

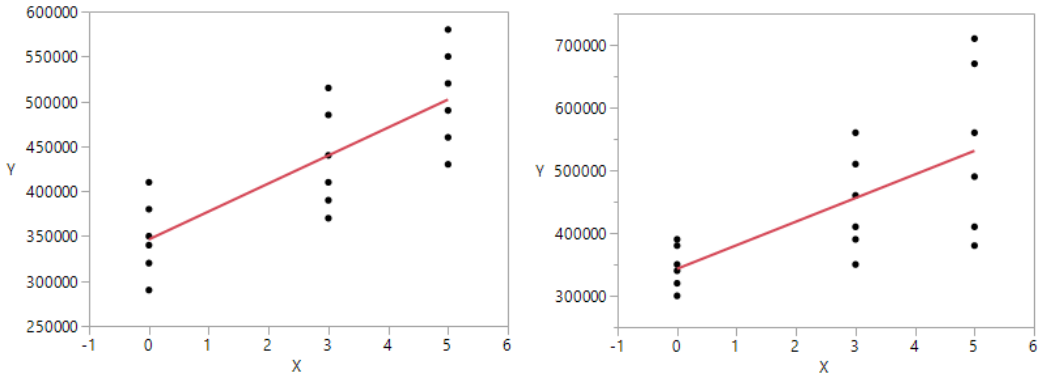
Assumption: the errors are homoskedastic (the assumption of constant variance)

Heteroskedasticity describes a phenomenon in which the error variance in the model is not constant. For instance, one might observe that the error variance depends on the predictors.

Consider the simple regression model

$$Y = B_0 + B_1 X + \varepsilon$$

The first figure below illustrates homoskedastic errors. Note how the variation around the regression line stays the same for all values of X. In other words, the variance appears in this case to be constant. The second figure illustrates heteroskedastic errors. In this case, it is evident that the variation (the spread around the regression line) increases with the value of X.



What are the consequences of heteroskedasticity?

In the presence of heteroskedasticity, the parameter estimates remain consistent and unbiased. However, the computed standard errors associated the estimates are biased. Thus, the usual *t* and *F*-statistic do not follow the t and *F*-distribution. Thus, any inference about the relationships in the model becomes invalid. The problem persists independently of sample size.

How to evaluate the assumption of homoskedasticity?

It is possible to test the assumption of homoskedastic errors. Consider the general regression model given by

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + \varepsilon$$

A test for heteroskedasticity is then performed using the following three steps:

1. Estimate the model and compute the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. Transform the residuals to obtain their squared values $\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2$.

2. Using the squared residuals, estimate a model of the form

   $$\varepsilon^2 = \gamma_0 + \gamma_1X_1 + \gamma_2X_2 + \cdots + \gamma_kX_k + \omega$$

   (estimating the model, one obviously has to use $\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2$ on the left hand side of the equation).

3. Evaluate the following hypothesis using the *F*-statistic

   $$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0 \text{ (homoskedasticity)}$$
   $$H_A : \text{at least one } \gamma_j \neq 0 \text{ for } j = 1, \ldots, k \text{ (heteroskedasticity)}$$

   Note that the test is a test for "overall significance", which simplifies the work somewhat since we can use the *F*-statistic from output of the software. If the null hypothesis is rejected, we conclude that the errors are heteroskedastic.

What to do in the presence of heteroskedasticity

Different strategies are available for handling the problem

- *Weighted estimation*: if the exact form of heteroskedasticity is known, it is possible to apply some form of weighting when the model is estimated. From a practical viewpoint, this strategy may not be very useful since we seldom know the exact form of heteroskedasticity.

- *Robust standard errors*: some software packages can provide robust standard errors. Note that robust standard errors are derived from large sample theory and may not be applicable in small samples.

- *Transforming the data*: sometimes heteroskedasticity is consequence of misspecification and wrong functional form (see previous lecture notes regarding functional form).

Assumption: there is no exact linear relationship among the independent variables $X_1, \dots, X_k$ *(no perfect collinearity)*

Perfect collinearity refers to the phenomenon that a predictor in a regression model can be written as a linear combination of the remaining predictors. More specifically, consider the general regression model

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + u$$

An exact linear relationship among the predictors is established if it possible to obtain a non-trivial solution of the equation

$$\lambda_1X_1 + \lambda_2X_2 + \cdots + \lambda_kX_k = 0$$

(a non-trivial solution is a solution where not all λs are simultaneously zero). Under this condition, it is not possible to uniquely determine the estimates of $B_0, \dots, B_k$ using OLS estimation. In practice, perfect collinearity is not a problem. What is more of a concern is the problem of *near perfect collinearity*. A useful measure of collinearity is the so-called *VIF* (Variance inflation factor) measure.

Assumption: the errors are normally distributed

In finite samples, a condition for the *t* and *F*-statistic to be *t* and *F*-distributed is that the error distribution is normal. However, due to the central limit theorem, as the sample grows larger, the *t* and *F*-statistic will approach (or approximate) the t and *F*-distribution regardless of the error distribution. Thus, when the number of observations in the data becomes large, the approximation to the *t* and *F*-distribution may be sufficient for reliable inference.