Data Problems

Henrique Veras

PIMES/UFPE

Intro

We now turn our attention to data problems.

The analysis to this point has assumed that the data in hand, $\mathbf{X}$ and $\mathbf{y}$, are well measured and correspond to the assumptions of the model and to the variables described by the underlying theory.

The cases we will examine are:

1. Multicollinearity
2. Missing values
3. Influential observations and outliers

Multicollinearity

One important thing to notice is that the Gauss-Markov theorem does not guarantee that the LS estimator has a small variance in any absolute sense.

We can write the expression for the conditional variance of $\mathbf{b}_k$ in the following way.

Define $\mathbf{x}_k$ as the column of the matrix $\mathbf{X}$ associated with the data from variable k.

Moreover, define the matrix $\mathbf{X}_{(k)}$ composed of the remaining columns of $\mathbf{X}$.

Multicollinearity

Thus, we can write

$$Var[b_k|\mathbf{x}] = \frac{\sigma^2}{(1 - R_k^2)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}$$

The factors that determine the precision of the kth LS coefficient estimator are:

1. R_k^2: greater R_k^2 increases $Var[b_k|\mathbf{x}]$ due to multicollinearity;
2. Variation in $\mathbf{x}_k$: higher $\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2$ lowers $Var[b_k|\mathbf{x}]$;
3. σ^2: greater σ^2 (poorer overall fit) increases $Var[b_k|\mathbf{x}]$.

When is multicollinearity a problem?

Some computer packages report a variance inflation factor (VIF): $1/(1 - R_k^2)$.

This is the increase in $Var[b_k|\mathbf{x}]$ that can be attributed to the correlation between the variables of the model.

The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

What to do when multicollinearity is a problem?

1. Get more data
   - But if more data is available, why would the analyst not use them to begin with?
2. Drop variables causing the problem
   - This creates a dillema: bias vs. precision.

The multicolinearity dillema

Consider the partitioned multiple regression

$$\mathbf{y} = \mathbf{X}\beta + z\gamma + \varepsilon$$

If we regress $\mathbf{y}$ on $\mathbf{X}$ only, the estimator is biased. Can you show this?

Moreover, we can show that

$$Var[\mathbf{b}|\mathbf{X}] \leq Var[\mathbf{b}_{\mathbf{X}.\mathbf{z}}|\mathbf{X}$$

Thus, we have a tradeoff: omitting a variable from the model improves precision but introduces bias.

Missing values and data imputation

When is it likely that some data is missing?

1. Surveys: some people might fail to answer questions.
2. Time series: some data measured at different frequencies;
3. Panel data: attrition

Missing values and data imputation

Cases to consider when data is missing:

1. Missing completely at random (MCAR): Affects efficiency but does not introduce any sort of bias on the estimated coefficients
2. Not missing at random (NCAR): Highly problematic, as it might introduce bias

Data imputation methods

To improve efficiency, we can input some information on the missing observations:

1. Zero-order method: Replacing missing x with $\bar{x}$ is equivalent to dropping the incomplete data.
2. Use the regression model to input the data.

It is important to notice, however, that all methods introduce some sort of *measurement error*, which causes bias (more on measurement error and what to do later on the course!)

Influential Observations

Given that the LS method is based on the *squared* deviations, the estimation is likely to be influenced by **extreme** observations.

An *influential* observation is one that is likely to have a substantial impact on the LS coefficients.

We can define an influence measure for observation $\mathbf{x}_i$ as

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_{(i)})^2}{\sum_{j \neq i}^{n} (x_j - \bar{x}_{(i)})^2}$$

We can say an observation is influential if $h_i > 2/n$.

Notice that the analysis is conditional on x_i only.

Influential Observations

What happens to the linear regression coefficient vector in a multiple regression when one observation is added to the sample?

We can show that the change in the linear coefficient is

$$\mathbf{b} - \mathbf{b_{(i)}} = \Delta\mathbf{b} = \frac{1}{1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i(\mathbf{y}_i - \mathbf{x}_i'\mathbf{b_{(i)}})$$

$\mathbf{b}$ is computed with observation i and $\mathbf{b_{(i)}}$ is computed without it.

An influential measure that is used is

$$h_{ii} = mathbfx_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i$$

The selection criterion would be $h_{ii} > 2(K-1)/n$.

Outliers

Outliers are observations that seem to come from outside the DGP/

Some reasons for outliers on the data:

1. Data error
2. Unusual residuals
3. Observation coming from a different population

To test for this latter case, we can construct *studentized residuals* by computing the regression coefficients and the residual variance without observation i for each observation in the sample and then standardizing the modified residuals.

Outliers

The ith studentized residual is

$$e(i) = \frac{e_i / \sqrt{1 - h_{ii}}}{\sqrt{\frac{\mathbf{e}'\mathbf{e} - e_i^2(1 - h_{ii})}{n - 1 - K}}}$$

Observations with $e(i) > 2$ would be considered outliers.

Using this procedure might be problematic as it should raise skepticism about the model specification in case a substantial proportion of the observations are considered outliers.

Outliers and Influential Points Graphically

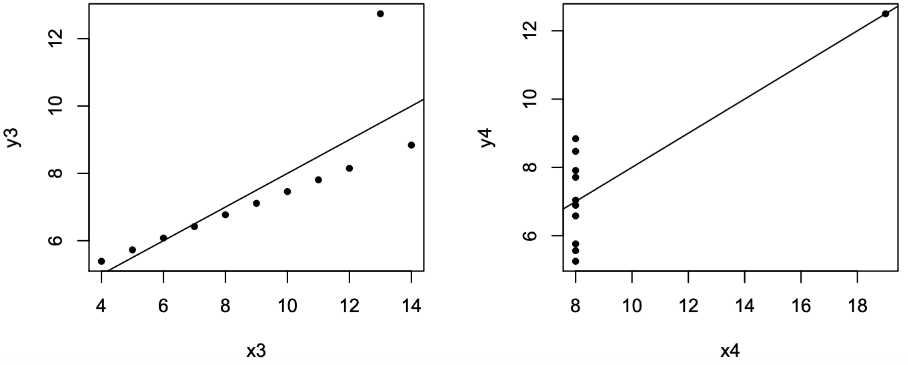

Table of Contents