9.5 Missing Data, Nonrandom Samples, and Outlying Observations

The idea of this section is to understand cases in which you have data which you suspect to have systematic biases i.e. the data is not random. In such cases, the OLS estimates will be biased and inconsistent.

9.5.1 Missing Data

1. Missing data: When data are missing for an observation on either the dependent variable or the independent variable, that particular observation cannot be used in multiple regression analysis.

   Is this a problem? The answer is no, if the data are missing completely at random (**MCAR**). The idea is that there is no systematic bias on the x's or the y which creeps into the data set due to missing observation.

   This implies that the remaining observations also form a completely random sample and dropping observations with missing values does not bias the randomness of the sample.

2. Missing indicator method (**MIM**): Let us assume that x_k has missing values, but we suspect that they are randomly missing. We want to use the information available from the observations for other x's. We do the following:

   (a) Create a variable $Z_{ik} = x_k$ when x_{ik} is observed, and zero otherwise.

   (b) Create another variable called m_{ik} which is equal to one when x_{ik} has a missing observation and 0 otherwise.

   (c) Run the following regression:

$$y_i \text{ on } x_{i1}, x_{i2}, \ldots x_{i,k-1}, Z_{ik}, m_{ik} \qquad \forall i$$

   Note: This method has good statistical properties only under strong assumptions. [Check for references in text.]

   Complete cases estimator is consistent even when the data are not randomly missing.

9.5.2 Non Random Samples

1. Exogenous sample selection: If the sample can be chosen on the basis of independent variables, it does not cause statistical problems and is known as exogenous sample selection.

2. This is often termed as missing at random (**MAR**), and requires that missingness in sample is unrelated to u but allows it to depend on $(x_1, x_2, \ldots x_k)$ whereas MCAR needs no relation between missing observations and $(x_1, x_2, \ldots x_k, u)$.

   (a) Consider a model in which we examine factors which determine savings. We conduct a survey of people over 35. This gives us non-random sample.

   (b) The argument here is that even this gives us unbiased conditional expectation function of a subset of population, provided there is enough variation in the independent variables in the sub-population.

3. Endogenous sample selection: If the sample is based on whether the dependent variable is above or below a given value, bias always occurs in OLS in estimating the population model. Such a selection is an example of *endogenous sample selection*.

   (a) For instance: if we are trying to understand wealth as a function of education, age and experience and we sample people with net wealth below $250000, we will witness biased and inconsistent estimators of the population parameter.

4. Non random sampling schemes could also lead to bias and inconsistency in estimators.

   (a) Stratified sampling - in which the population is divided into nonoverlapping, exhaustive groups, or strata. Then, some groups are sampled more frequently than is dictated by their population representation, and some groups are sampled less frequently.

   (b) Check out the explanation in the text about pay for women in military.

5. Classic selection bias example: Suppose we want to understand the potential wage earnings as a function of education. We collect a sample of workers and their education. However, we will potentially run into biased estimates because we might systematically collect non-random sample since those who do not earn a wage might be systematically less educated. Thus the dependent variable is non-random.

9.5.3 Outliers and Influential Observation

1. Influential observations: an observation is an influential observation if dropping it from the analysis changes the key OLS estimates by a practically "large" amount.

2. Outliers: also cause bias as OLS relies on minimizing sum of squared residuals and outliers would generate large residuals.

   - When outliers are introduced due to manual error in data entry etc. there is no simple way to deal with them. Sometimes they appear in simple box plots and other times, they are more difficult to detect.
   - Outliers could arise when sampling from small population, in which case there is no apriori method of handling it.
   - Outliers should not be identified by the size of the residual under OLS, because OLS estimates are based on minimizing these residuals. So choosing outliers on the basis of OLS residual results in circular logic.

3. Detecting outliers using Studentized residuals: For the observation h which is the outlier in your data, define a dummy variable equal to one for that observation and include the dummy variable in the regression.

   - ▷ The coefficient on the dummy variable has a useful interpretation: it is the residual for observation h computed from the regression line using only the other observations. Therefore, the dummy's coefficient can be used to see how far off the observation is from the regression line obtained without using that observation.
   - ▷ The t-stat on dummy variable has an t_{n-k-2} distribution and a large value implies a large residual relative to its estimated standard deviation.
   - ▷ Note: t-stat cannot tell you how important the observation is for calculation of β's. Whether dropping one observation significantly changes the β needs to be calculated independently.
   - ▷ Certain functional forms are less sensitive to outliers and certain observations stare at you proclaiming to be the outlier.

9.6 Least Absolute Deviations Estimation

1. Another estimation method which does not get affected by outliers is the Least Absolute Deviations (LAD) which obtains b_k by solving the following:

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} \left| y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik} \right|$$

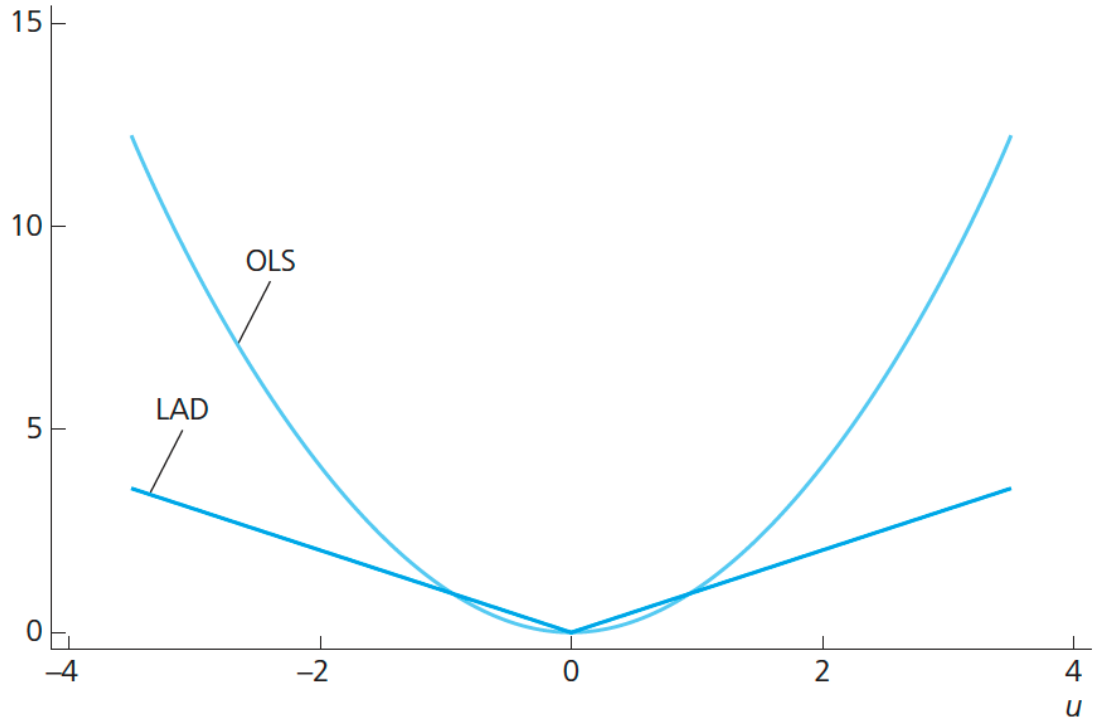
We cannot express the b_k in terms of the data since there is no closed form solution possible for them.

Interpretation: it is known that LAD is designed to estimate the parameters of the conditional median of y given $x_1, x_2, \ldots, x_k$ rather than the conditional mean. A limiting form of quantile regression. Because the median is not affected by large changes in the extreme observations, it follows that the LAD parameter estimates are more resilient to outlying observations.

2. LAD is a robust regression which technically means that it is insensitive to extreme observations.

3. Notice the objective functions of OLS and LAD estimation procedures. The LAD is linear on either side of zero while OLS gives increasing importance to large residuals making them sensitive to outliers.

FIGURE 9.2 The OLS and LAD objective functions.



4. LAD estimators only have an asymptotic justification.

5. A more subtle but important drawback to LAD is that it does not always consistently estimate the parameters appearing in the conditional mean function, $E(y|x_1, x_2, \ldots, x_k)$.

   - ▷ OLS produces unbiased and consistent estimators of the parameters in the conditional mean function whether or not the underlying error distribution is symmetric or not.
   - ▷ Under the conditional independence of population error u on $(x_1, x_2, \ldots, x_k)$, estimates of LAD and OLS should differ only by how much mean and median differs. However, the conditional independence of u_{LAD} is stronger than u_{OLS}.

6. LAD allows us to obtain easy partial effects using monotonic transformations. For instance, in the log dependent variable model we have the conditional median as

$$Med(y|x) = exp(\beta_0 + \mathbf{x}\boldsymbol{\beta}) \qquad \text{when} \;\; Med(u|\mathbf{x}) = 0$$

   We need not assume that u and $\mathbf{x}$ are independent and that the conditional median holds for any distribution of u.