9 More on Specification and Data Issues

1. Note: contents in Chapter 8 are only at the stage of inference. As mentioned previously, an issue of first order importance is the issue of specification which causes breakdown of conditional independence. In this chapter we will deal with 3 issues which could cause such a breakdown

   (a) Functional form mis-specification: We will examine ways to test functional form mis-specification and be able to detect problems with our model using such tests.

   (b) Omitted Variable Bias: could be in part dealt with by including proxy variables.

   (c) Measurement errors: could also induce bias in our estimates.

9.1 Functional Form Misspecification

1. Consider the true relation between wage and experience in the population as follows:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

2. When we estimate the model without $exper^2$ instead, we commit a misspecification. We bias β_1, our key parameter of interest measuring "returns to education". Size of bias depends on:

   (a) size of β_3 in the population

   (b) Correlation between edu, $exper$ and $exper^2$

3. Another issue in misspecification: We cannot interpret β_2 as the correct partial effect of experience.

4. Misspecifying dependent variable form could also lead to unbiased or inconsistent estimates.

5. Broad point: functional form is a mere trial and error issue especially when you have all kinds of data to accommodate different functional forms

6. Economic model of crime: Economists have been trying to understand crime and criminal activity in the last 3 decades. They use regression anlaysis to understand the factors which increase the probability of committing a crime. The dependent variable is number of arrests in 1986 ($narr86$) which is explained by $pcnv$ i.e. percentage conviction, average sentence ($avgsen$), prison time last year ($ptime86$) etc.

   (a) Consider example 9.1. Get the data set from JW's online companion and try and run the two models (try different functional forms: with interaction terms)

   (b) Check whether they are jointly significant to test nested models using F-stat. Note that the presence of interactions and quardratics makes for difficult interpretation.

   (c) The quardratic terms on $pcnv$ is significant and implies that previous conviction increases chance of conducting a crime initially, but then acts as a deterrent. This becomes difficult to interpret.

9.1.1 Ramsey's RESET Test for Functional Form

1. Manually adding certain quardratics and functional form specification might not help because:

   (a) It uses degrees of freedom and might not detect the functional form exactly.

   (b) It neglects other kinds of non-linearities.

2. RESET does something which we have done in previously. From the basic linear model obtain predictions

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

   Then run the following model examining quardratics and cubes of independent variables:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + v$$

3. RESET is an F-statistic for testing $H_0 : \delta_1 = 0, \delta_2 = 0$ in the expanded model. If we reject the null, then we have a functional form problem. The resultant F-stat is approximately $F_{2,n-k-3}$ in large samples under MLR-1 to MLR.6 assumptions.

4. Caution:

   (a) RESET test does not tell you what to do after you reject the null.

   (b) RESET only tests functional form and not whether you've actually got an unbaised and consistent estimate. That is the hard job you would have to do on your own!

9.1.2 Non Nested Alternatives

1. Pseudo model approach: in which you nest two non-nested models under a larger model and text exclusion restrictions using standard F-test i.e.

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 log(x_1) + \gamma_4 log(x_2) + u$$

   The two models are ones with level and other with log on x_1 and x_2. The null hypothesis would be $H_0 : \gamma_3 = 0, \gamma_4 = 0$.

2. David-MacKinnon's approach: suggested that if the conditional expectation function for y is $E_1(y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ then the fitted values from the log model $E_2(y|x) = \beta_0 + \beta_1 log(x_1) + \ldots + \beta_k log(x_k) + v$ should have insignificant coefficients when added to E_1. We just examine the t-statistic for the primary model to check the hypothesis.

3. Problems with Non-nested Testing:

   (a) At times there are no clear winners between two competing models.

   (b) Rejecting a model does not point us to the correct model. That still needs guesses!

9.2 Using Proxy Variables for Unobserved Explanatory Variables

1. Consider the following model for wages

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 ability + u$$

Our problem is that we know *ability* affects wages, but have no way to know how. Non-inclusion of ability biases our estimate on education and experience – which is what we care about.

2. Using proxy variables: A proxy variable is something that is related to the unobserved variable that we would like to control for in our analysis. For instance, let us assume that IQ scores are a proxy for the unobservable *ability*.

3. Thinking about ability and IQ: we need them to have some relationship and we do not need ability to be completely explained by IQ. This implies:

$$ability = \delta_0 + \delta_1 IQ + \delta_2 somethingelse + v$$

4. The requirement for a proxy variable for ability is that δ_1 $neq0$. Then we use IQ instead of ability in our original regression model to obtain a **plug-in solution to the omitted variable problem**. However, there are two conditions under which this solution could lead to consistent estimators of β_1 and β_2.

   (a) The error term should be uncorrelated with *educ*, *exper*, *ability* and IQ, i.e. the omitted variable as well as the proxy variable.

   (b) The error term v should be uncorrelated with *educ*, *exper* and IQ

$$E(abil|educ, exper, IQ) = E(abil|IQ) = \delta_0 + \delta_1 IQ$$

5. Note: you could do the algebra on a general model with x_o as omitted variable and x_p as proxy variable. You can substitute predicted x_o i.e. $\hat{x}_o = \hat{\delta}_0 + \hat{\delta}_1 x_p$ into the main model and obtain the modified Gauss Markov assumptions in the augmented model.

6. Consider the returns to education i.e. the $\hat{\beta}$ on *educ* without proxying for ability. The estimated returns to education is 6.5% (example 9.3 and computer exercise). When JW includes IQ scores in this model, then the returns fall to 5.4%. This tells us that by correcting for ability bias, one could find the true returns to education.

7. Multicollinearity vs Appropriate proxy variable: It is possible that there could be potential multicollinearity between IQ and education. This could make our estimates imprecise.

   (a) But including IQ leads to lower error variance and more precise estimates.

   (b) Also, including IQ at the cost of some multi-collinearity might be required to get a more unbaised estimator.

9.2.1 Using Lagged Dependent Variables as Proxy Variables

1. Idea: If we have no proxy variables which we could think of, we could alternatively use lagged dependent variables with the idea that it would proxy a part of the unobservable omitted variable.

2. Example: consider the model on city crime rates:

$$crime = \beta_0 + \beta_1 unem + \beta_2 expend + \beta_3 crime_{t-1} + u$$

where $unem$ is the unemployment rate in the city, $expend$ is the expenditure on law and enforcement and $crime_{t-1}$ is the previous year's per capita crime rate.

The idea of $crime_{t-1}$ is to capture this idea that cities which have higher crime rate historically would have spend more on law and enforcement hence leading to lower crime. Thus factors unobserved by us that affect current $crime$ could be correlated with $expend$ or $unem$. By including $crime_{t-1}$, we take part of the stuff from u to the equation.

9.2.2 A Different Slant on Multiple Regression

1. Tired of thinking about ideal relationship between y and $x_1, x_2, \ldots, x_k$?

2. Instead: deal with y on a subset of $x_1, x_2, \ldots, x_k$ which are observable and try to focus on getting a consistent and unbiased estimate conditional on the observables.

3. Does this solve the bias induced by leaving out the unobservable?

4. No, it does not. But if it is the next best thing possible, then go with what you can with the data you have rather than thinking about what you ought to have!

   My take: This is a practical empirical solution, but it does not help us completely. I would still run behind the understanding the direction of bias caused by unobservables.

9.2.3 Potential Outcomes and Proxy Variables

The conclusions in this section come from ATE dealt with in section 7.6.

1. Consider a binary treatment indicator w and $y(0)$ and $y(1)$ denote without treatment and with treatment predicted values.

2. In a scenario when we have unobservables $v(0)$ and $v(1)$, we could proxy it with $\mathbf{x} = (x_1, x_2, \ldots, x_k)$.

3. This implies that conditional on the $\mathbf{x}$, our program w is independent and tells us about its effect on the dependent variable of interest.

4. This takes the following form:

$$y(0) = \mu_0 + v(0) \qquad y(1) = \mu_1 + v(1)$$
$$E[v(0)|w, \mathbf{x}] = E[v(0)|\mathbf{x}] = (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\beta_0}$$
$$E[v(1)|w, \mathbf{x}] = E[v(1)|\mathbf{x}] = (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\beta_1}$$

9.3 Models with Random Slopes

1. Idea: Remember the model where dummy variable interacted with another continuous variable? Similarly, what if partial effect of a variable depends on unobserved factors that vary by population unit as follows:

$$y_i = a_i + b_i x_i$$

In our usual model we have $E(b_i) = \beta_1$ and $a_i = \beta_0 + u_i$. This model is called the **random coefficient model** or random slope model.

The unobserved slope coefficient b_i is considered a random draw from the population along with observed data (x_i, y_i). So in some sense, for every individual i, you know his/her x_i, y_i and b_i.

2. *Estimation*: In such cases we estimate the average response across individuals denoted by $E(b_i) = \beta_1$ known as the *average partial effect* (APE) or average marginal effect.

3. To understand these models as versions of our usual model we write them as follows:

$$y_i = a_i + b_i x_i = \beta_0 + c_i + \beta_1 x_i + d_i x_i = \beta_0 + \beta_1 x_i + u_i$$

This tells us that we could potentially write the random part of the model into its two components (a) the average partial effect and (b) part of the random effect. For instance $a_i = \beta_0 + c_i$.

(a) Conditional independence assumption implies $E(u_i|x_i) = 0$. Both the models produce unbiased estimates.

(b) The error term i.e. u_i contains heteroskedasticity. If $Var(c_i|x_i) = \sigma_c^2$, $Var(d_i|x_i) = \sigma_d^2$ and $Cov(c_i, d_i|x_i) = 0$, then

$$Var(u_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$$

This could be addressed using methods learnt in chapter 8. Notice that random slope models could be sources of heteroskedasticity. Caution: this does not allow for heteroskedasticity in a_i or b_i. This implies that we cannot distinguish between a random slope model, where the intercept and slope are independent of x_i, and a constant slope model with hereoskedasticity in a_i

4. **sSummary**: Estimating random slope models as used in chapter 6 is easy if slopes are independent of the explanatory variables.

9.4 Properties of OLS under Measurement Error

1. Contrasting the case of omitted variable with measurement error: In the measurement error case, the variable that we do not observe has a well-defined, quantitative meaning (such as a marginal tax rate or annual income), but our recorded measures of it may contain error.

   In certain cases we could also calculate the size of the error asymptotically.

   - ▷ Thus econometrician has to use annual income of individuals to explain the dependent variable, but she does not have the data on annual incomes. She goes and asks people their annual incomes. Because people do not report their true income, the data collected is only an approximation of the actual annual incomes.

2. In measurement error problem, the mis-measured variable is of interest, whereas in the case of OVB, we might just be interested in its effect on the other key variables of interest.

9.4.1 Measurement Error in the Dependent Variable

1. Consider the following model in which we would ideally like to have data on y^*, but have an observable measure of y. For instance it is hard to obtain data on savings i.e y^* instead families reports a measure of savings y (self-reported measure):

$$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

2. Measurement error is defined as the difference between the observed value and the actual value

$$e_0 = y^* - y$$

If we substitute y^* in the equation with measurement error, we get a new composite error of the following form:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0$$

If e_0 has zero mean, then we get unbiased estimates. Even if e_0 has a non-zero mean, we do not worry about the bias as much.

3. Multiplicative measurement error: what if we have a measurement error of the following form: $y = y^* a_0$. In such a scenario, we take log on both sides and obtain $log(y) = log(y^*) + log(a_0)$, where $e_0 = log(a_0)$.

4. Conditional independence of the measurement error: If e_0 and x_j are statistically independent, then our estimates are consistent and unbiased. Our usual OLS inference procedures apply.

9.4.2 Measurement Error in the Explanatory Variable

1. Consider the simple regression model in which we have the desired independent variable which we want to explain the dependent variable:

$$y = \beta_0 + \beta_1 x_1^* + u$$

But instead of x_1^* we have a variable x_1 which is ridden with measurement error. We assume that the average measurement error is zero, which is not crucial. We assume more importantly that $E(y|x_1^*, x_1) = E(y|x_1^*)$.

2. Measurement error and independent variable: If we have the observed variable uncorrelated with the measurement error i.e.

$$Cov(x_1, e_1) = 0$$

we could simply plug in the observed variable and run the regression without affecting the properties of the OLS estimator. We estimate the following equation:

$$y = \beta_0 + \beta_1 x_1 + (u + \beta_1 e_1)$$

3. The estimates are consistent. Error variance is

$$Var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$$

4. If instead the covariance between the unobserved variable with the measurement error takes the following form:

$$Cov(x_1^*, e_1) = 0$$

then the assumption is refered to as the *Classical error in variable* assumption. However, if this holds, then the earlier assumption cannot be true i.e. $Cov(x_1, e_1) \neq 0$. Instead the correlation between the two is

$$Cov(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1^2) = 0 + \sigma_{e_1}^2$$

5. This implies that our regression in point (2) does not satisfy the CLRM assumptions. The covariance between the x_1 and the composite error $u - \beta_1 e_1$ is

$$Cov(x_1, u - \beta_1 e_1) = -\beta_1 Cov(x_1, e_1) = -\beta_1 \sigma_{e_1}^2$$

Thus the OLS regression of y on x_1 gives a biased and inconsistent estimator.This implies the following:

$$plim(\hat{\beta}_1) = \beta_1 \overbrace{\left(\frac{\sigma_{x_1}^{2*}}{\sigma_{x_1}^{2*} + \sigma_{e_1}^2} \right)}^{<1}$$

This implies that for a positive β_1, under CEV assumptions, the estimated $\hat{\beta}_1$ will be smaller than its population counterpart, β_1. This is called the "attenuation bias" in OLS.

6. In a multiple regression model with measurement error for x_1^*, under the CEV assumptions, OLS will be biased and inconsistent. However, in this case all OLS estimatoes will be biased and not just the ones associated with x_1 - the measure of x_1^* in the sample for the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u - \beta_1 e_1$$

The attenuation bias can now be obtained as:

$$plim(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{r_1}^{2*}}{\sigma_{r_1}^{2*} + \sigma_{e_1}^2} \right)$$

where r_1^* is the population error in the equation $x_1^* = \alpha_0 + \alpha_2 x_2 + \alpha_3 x_3 + r_1^*$.

We cannot say anything apriori and clearly about the biases in estimating β_2 and β_3 in this model.

7. Summing up: Generally the CEV assumption is still a strong assumption. It is more likely that the measurement error, e_1 is likely to be correlated with both x_1 and x_1^*.

   (a) IV estimation could provide us consistent estimators even in the presence of a general measurement error problem.

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:

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

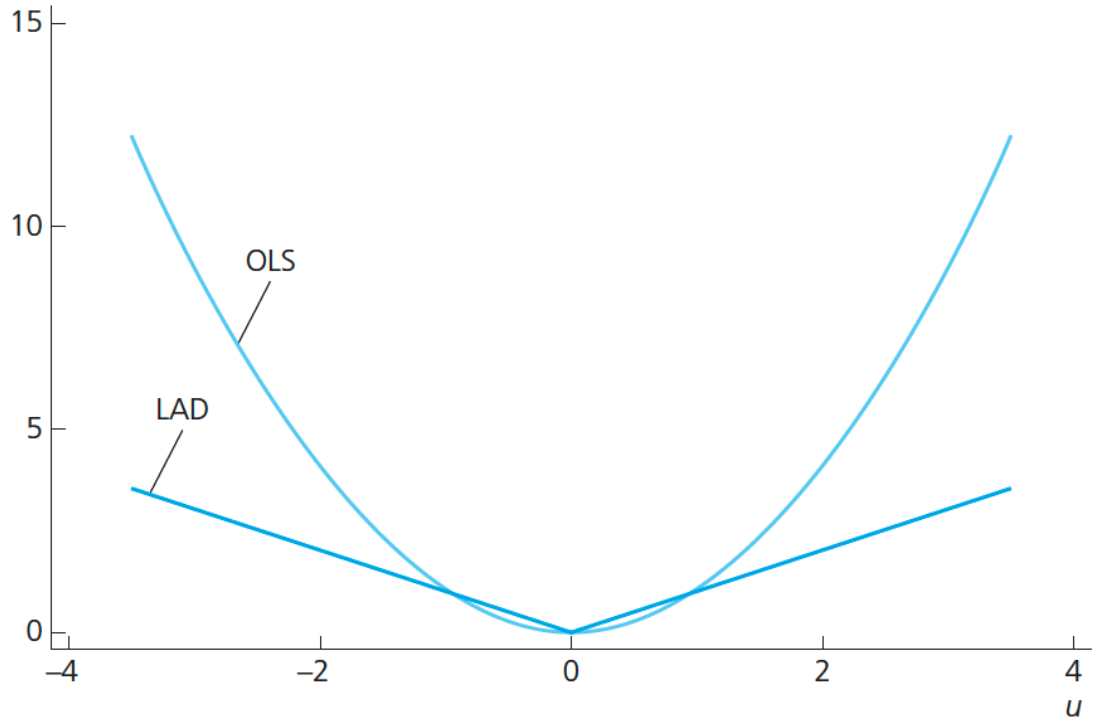
We cannot express the b_k in terms of the data since there is no closed form solution possible for them.

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.