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!