More on Specification and Data Issues

- Heteroskedasticity problem - not serious

- Specification error problem -serious!

1 Functional Form Misspecification

Suppose the true model is

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

- But we do not include $\exp er^2$, and estimate

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + u.$$

- Our $\beta_0, \beta_1, \beta_2$ could be biased.
- The severity of biasedness depends on

1.1 The RESET test

Rationale behind – add polynomial terms $(x^2, x^3, ...)$ to the model and test whether the coefficients are significant. If significant, then we need to include non-linear terms.

Suppose we have a linear baseline model to begin with

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \tag{11.1}$$

To save the # of regressors (save d.f.), we can instead estimate

1.2 Tests against Nonnested Alternatives

For example, to test

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad \text{(Model \#1)}$$

against

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u. \qquad \text{(Model \#2)}$$

Which one should we use? .... There are many methods which can help us justify:

1. Use the RESET test

2. Mizon and Richard (1986)

3. Davidson-MacKinnon Test

Rationale - If Model#1 is true, then the fitted values ($\hat{y}$) from Model#2 should not be significant!

2 Using Proxy Variables for Unobserved Explanatory Variables

In many cases, we cannot observe some important variables

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 ability + u.$$

- Ability is unobserved, but is very likely to explain wage or log(wage).

- Ability is very likely to be correlated with *educ* and $\exp er$. If ability is omitted, $\hat{\beta}_1$ and $\hat{\beta}_2$ will be biased.

Remedy (use proxy variables (this chapter) or use instrumental variable method (not in this chapter))

2.1 Using Proxy variables

2.2 Using Lagged Dependent Variables as Proxy Variables

3 OLS with Measurement Error

Suppose

$$GDP = \beta_0 + \beta_1 \exp ort + \beta_2 investment + \beta_3 \inf lation + \beta_4 educ + ... + u$$

- How do you know if a country's GDP is correctly measured?
- What about the measurement of other variables?

3.1 Measurement Error in the Dependent Variable

If there is a measurement error in y variable (dependent variable), then

3.2 Measurement Error in the Explanatory Variable

Suppose the true model is

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

4 Missing Data, Nonrandom Samples, Outlying Observations

4.1 Missing Data

Household id.	head_educ	total_income	members
1	12	12,000	2
2	14	25,000	3
3	.	60,000	2
4	4	.	4
5	8	120,000	4
6	16	32,900	2

- STATA would drop the observation with missing data. So, household 3 and 4 would not be used in the regression.
- **If the missing data are missing at random, then $\hat{\beta}_{OLS}$ would not be biased.**

4.2 Nonrandom Samples

4.3 Outliers and Influential Observations

5 Models with Random Slopes

- What if the partial effects of a variale depends on unobserved factors that "vary" by population unit?

$$y_i = a_i + b_i x_i$$

- This is called a random coefficient model, or a random slope model (u_i would then be absorbed into a_i).
- a_i, b_i are viewed as a random draw from the population along with the observed data.
- In any case, we cannot estimate the actual a_i, b_i (for each population unit). We can only find their average (average partial effect, APE).

$$\beta_0 = E(a_i) \quad and \quad \beta_1 = E(b_i).$$