AFRE 835: Introductory Econometrics

Chapter 9: Miscellaneous Issues

Spring 2017

Introduction

- This chapter touches on a series of additional issues, mostly centering around functional form misspecifications.
- The chapter also discusses the impact of
  - measurement errors in the dependent and independent variables;
  - missing data;
  - departures from random sampling.

Outline

Functional Form Misspecification

- Functional form misspecification arises if we have specified the incorrect population regression function, linking the dependent variable of interest to the observed independent variables.
- Two common concerns are:
  - Insufficiently flexible functional form (e.g., excluding quadratic or higher terms)
  - The wrong form for the dependent variable (e.g., levels versus logs);

Log(wage)

- Consider a model of wages (using WAGE1.DTA), where we specify

$$E[ln(wages)|educ, exper, female) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female.$$

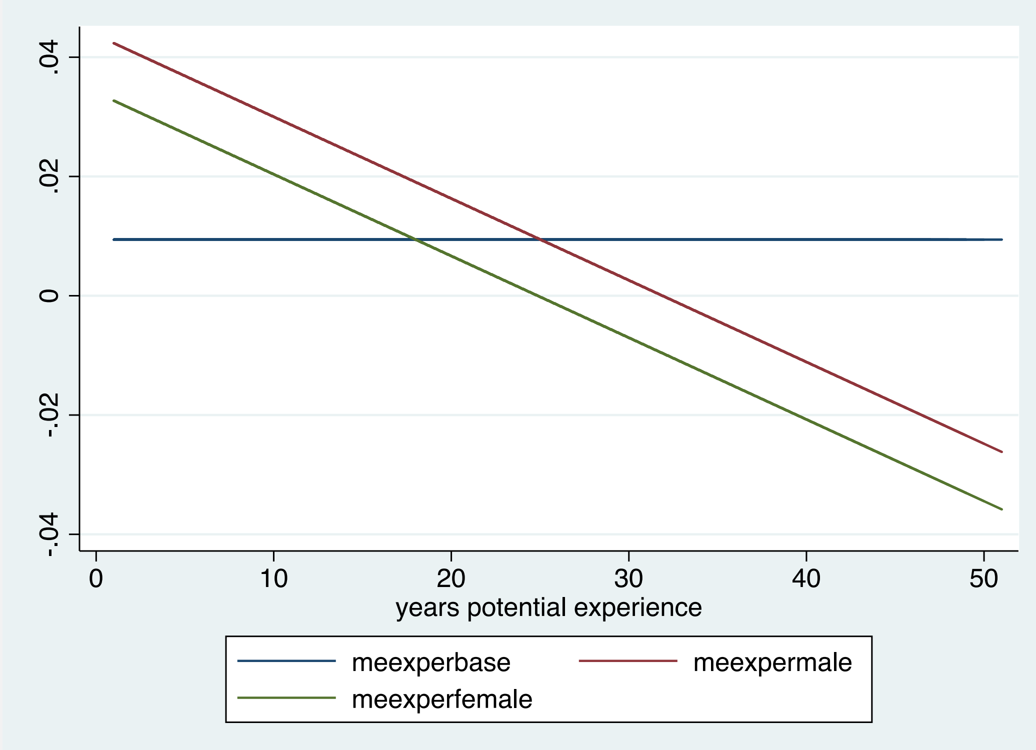
- This assumes that the marginal effect of experience and education on log-wages is constant.
- It also ignores potential differential returns to either education or experience by gender.
- A more general functional form would include quadratic terms in experience and/or educ and interaction effects by gender.
- For example, consider the specification focused on possible nonlinear and interaction effects of experience

$$E[ln(wages)|educ, exper, female) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female + \beta_4 * exper^2 + \beta_5 * exper * female$$

log(wages)

	basic	nonlinear
educ	0.091	0.087
	(0.007)**	(0.007)**
exper	0.009	0.044
	(0.001)**	(0.005)**
female	-0.344	-0.172
	(0.038)**	(0.058)**
expersq		-0.001
		(0.000)**
femexper		-0.010
		(0.003)**
_cons	0.481	0.273
	(0.105)**	(0.106)*
R^2	0.35	0.41
N	526	526

* $p < 0.05$; ** $p < 0.01$



Testing for Functional Form Misspecification

- One approach to considering functional form misspecifications is to include quadratic and interaction terms into the model and use standard $F-$statistics to test for their joint significance.
- In the previous example, the F-statistic corresponding to the null hypothesis $H_0 : \beta_4 = \beta_5 = 0$ would be 27.7, which is greater the $F_{2,522} = 4.61$ critical value using a 1% significance level.
- However, incorporating a full set of quadratic and interaction terms can *use up* a large number of degrees of freedom if k is large.
- The **regression specification error test (RESET)** suggests a more parsimonious test by
  - Fitting a simple linear model and recovering fitted values $\hat{y}$.
  - Estimate an expanded model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error \qquad (1)$$

  testing the null hypothesis $H_0 : \delta_1 = \delta_2 = 0$. Not commonly used.

Testing Against Non-nested Alternatives

- One is often faced with a choice between *non-nested* competing models; e.g.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{2}$$

vs.

$$y = \beta_0 + \beta_1 ln(x_1) + \beta_2 ln(x_2) + u \tag{3}$$

- One approach is to specify a more general model, with

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 ln(x_1) + \beta_4 ln(x_2) + u \tag{4}$$

  and test $H_0 : \beta_1 = \beta_2 = 0$ and $\tilde{H}_0 : \beta_3 = \beta_4 = 0$.
- The problem is you may end up rejecting both or neither specifications.
- Moreover, there are other possibilities; e.g., $\check{H}_0 : \beta_1 = \beta_4 = 0$.

The Davidson-MacKinnon Test

- The approach on the previous slide can be problematic if k is large.
- Davidson and MacKinnon suggested a more parsimonious test by, first, estimating the two competing models:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{5}$$

vs.

$$y = \beta_0 + \beta_1 ln(x_1) + \cdots + \beta_k ln(x_k) + u \tag{6}$$

  ...and then estimating

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \theta_1 \hat{\hat{y}} + u \tag{7}$$

vs.

$$y = \beta_0 + \beta_1 ln(x_1) + \cdots + \beta_k ln(x_k) + \theta_1 \hat{y} + u \tag{8}$$

  where $\hat{y}$ and $\hat{\hat{y}}$ are fitted values from (5) and (6), respectively, and testing $H_0 : \theta_1 = 0$

The Problem of Unobserved Explanatory Variables

- A common problem arising in applied research is that we may not have the precise variable we need for our analysis.
- The classic example here emerges in the labor literature, in models designed to assess the returns to education.
- One might assume that the following relationship holds

$$ln(wages) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u \tag{9}$$

  where *abil* measures the underlying ability of the individual.
- Unfortunately, ability is a difficult characteristic to measure.
- Excluding it from the model, however, is not a good option, since ability would then be absorbed by the error term

  ...and, since ability is almost surely correlated with *educ*, we would end up with omitted variables bias in estimating the returns to education.

The Use of Proxy Variables

- One solution to the unobserved variable problem is to employ a **proxy variable** - a variable related to the unobserved variable, but in a specific way.
- Consider a model with three independent variable (as in the log-wage model on the previous slide); i.e.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u \tag{10}$$

  where x_3^* is unobserved.
- Now suppose that we have a proxy variable for x_3^*, say x_3 where

$$E(x_3^*|x_1, x_2, x_3) = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 \tag{11}$$

  ...or simply

$$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3 \tag{12}$$

  with $E(v_3|x_1, x_2, x_3) = 0$.
- What happens if we simply use x_3 instead of x_3^*?

The Use of Proxy Variables (cont'd)

- In particular, suppose we estimate the model

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_3 x_3 + \tilde{u} \tag{13}$$

- Substituting in (12) into (10), we know that

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3) + u \\ &= (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \beta_3\delta_3 x_3 + \beta_3 v_3 + u \end{aligned} \tag{14}$$

- Comparing (13) and (14) we have that:

$$\tilde{\beta}_j = \beta_j + \beta_3\delta_j \quad j = 0, 1, 2 \tag{15}$$

$$\tilde{\beta}_3 = \beta_3\delta_3 \tag{16}$$

$$\tilde{u} = \beta_3 v_3 + u \tag{17}$$

The Use of Proxy Variables (cont'd)

- We need one additional assumption; i.e., u is uncorrelated with x_3.
- The zero condition mean assumption holds for (13), since

$$E(\tilde{u}|x_1, x_2, x_3) = \beta_3 E(v_3|x_1, x_2, x_3) + E(u|x_1, x_2, x_3) = 0 \tag{18}$$

- As a result, OLS applied to (13) will yield unbiased and consistent estimates of $\tilde{\beta}_j, j = 0, \ldots, 3$.
- The only issue that remains is, under what conditions can we use the OLS estimator to recover parameters of interest?
  - We need $\delta_3 \neq 0$, otherwise the unobservable variable is not an issue to begin with.
  - In order to recover β_1 and β_2, we need $\delta_1 = 0$ and $\delta_2 = 0$, in which case $\tilde{\beta}_j = \beta_j \; j = 1, 2$.
  - Typically, we don't care about β_0, but if we do, then we need $\delta_0 = 0$.
  - Identifying β_3 requires $\delta_3 = 1$, in which case we essentially are observing the so called unobserved variable.

Returns to Education Example

- In the returns to education example, ability is proxied for using IQ.
- Our key assumption is then

$$E(abil|educ, exper, IQ) = \delta_0 + \delta_3 IQ \tag{19}$$

...i.e., mean ability varies with IQ, but (conditional on IQ) , ability does not vary with education or experience.

Returns to Education Example

TABLE 9.2 Dependent Variable: log(*wage*)

Independent Variables	**(1)**	**(2)**	**(3)**
educ	.065 (.006)	.054 (.007)	.018 (.041)
exper	.014 (.003)	.014 (.003)	.014 (.003)
tenure	.012 (.002)	.011 (.002)	.011 (.002)
married	.199 (.039)	.200 (.039)	.201 (.039)
south	−.091 (.026)	−.080 (.026)	−.080 (.026)
urban	.184 (.027)	.182 (.027)	.184 (.027)
black	−.188 (.038)	−.143 (.039)	−.147 (.040)
IQ	—	.0036 (.0010)	−.0009 (.0052)
educ·IQ	—	—	.00034 (.00038)
intercept	5.395 (.113)	5.176 (.128)	5.648 (.546)
Observations *R*-squared	935 .253	935 .263	935 .263

Models with Random Slopes

- In our models thus far, we have assumed that the slope coefficients are the same for everyone.

  ...or at least that they only vary in observable ways.
- For example, in the model in Table 9.2 above, we assume that the marginal effect of *educ* on log-wages is given by:

$$\frac{\partial E[log(wages)|\mathbf{x}]}{\partial educ} = \beta_{educ} + \beta_{educ \cdot IQ} \cdot IQ \tag{20}$$

- For a given IQ, and holding everything else constant, the marginal impact of education is the same for everyone.
- But it might be the case that this marginal effect varies by individual because there are *unobserved factors* interacting with education that also impact log-wages.
- These interaction terms can lead to *random slope* (or *random coefficient*) models.

An Illustration

- Suppose that we have the following population model for our dependent variable of interest

$$\begin{aligned} y_i &= \alpha + \gamma z_i + \beta x_i + \delta x_i z_i + \tilde{u}_i && (21) \\ &= (\alpha + \gamma z_i) + (\beta + \delta z_i)x_i + \tilde{u}_i \\ &= (\alpha + c_i) + (\beta + d_i)x_i + \tilde{u}_i \\ &= a_i + b_i x_i + \tilde{u}_i && (22) \end{aligned}$$

  where $c_i = \gamma z_i$ and $d_i = \delta z_i$.
- Without loss of generality, assume that $E(z_i) = 0$.
- Think of $a_i = \alpha + \gamma z_i$ and $b_i = \beta + \delta z_i$ as representing the intercept and slope terms, respectively, for the relationship between y_i and x_i.
- If we observe x_i and z_i, we would simply estimate the model depicted in (21) as including an interaction term between x_i and z_i.

An Illustration (cont'd)

- Suppose, however, that z_i is unobserved.
- The model in (22) has intercept and slope terms that vary by individual, with $y_i = a_i + b_i x_i + \tilde{u}_i$.
- The individual specific slope term implies that the marginal effect of a change in x_i varies by individual.
- We cannot estimate a separate intercept and slope for each individual.
- The question is can we estimate the average intercept $(E[a_i] = \alpha)$ and average slope $(E[b_i] = \beta)$ in the population?
- β is referred to as the **average partial effect (APE)** or the **average marginal effect (AME)**.

An Illustration (cont'd)

- We can re-write our model as:

$$\begin{aligned} y_i &= (\alpha + c_i) + (\beta + d_i)x_i + \tilde{u}_i \\ &= \alpha + \beta x_i + (c_i + d_i x_i + \tilde{u}_i) \\ &= \alpha + \beta x_i + u_i \end{aligned} \tag{23}$$

  where $u_i = c_i + d_i x_i + \tilde{u}_i$.
- What do we need for OLS applied to the above equation to yield unbiased estimates of α and β?
- We need:

$$\begin{aligned} 0 = E(u_i|x_i) &= E(c_i + d_i x_i + \tilde{u}_i|x_i) \\ &= E(c_i|x_i) + E(d_i|x_i)x_i + E(\tilde{u}_i|x_i) \end{aligned} \tag{24}$$

- Sufficient conditions for this to hold are: $E(c_i|x_i) = 0$, $E(d_i|x_i) = 0$, and $E(\tilde{u}_i|x_i)$.

An Illustration (cont'd)

- These conditions essentially imply that the unobserved factors are mean independent of the observed regressors.
- One additional note: The structure of the error term implies that we have heteroskedasticity in this case.
- Specifically, if $Var(c_i|x_i) = \sigma_c^2$ and $Var(d_i|x_i) = \sigma_d^2$, then

$$Var(u_i) = \sigma_c^2 + \sigma_{\tilde{u}}^2 + \sigma_d^2 x_i^2 \tag{25}$$

- The above model can also be generalized to allow the intercept and slopes to vary in both observable and unobservable ways.

Measurement Error

- Not surprisingly, analysts do not always have perfect measures of the variables of interest to them.
- Measurement errors can arise due to flawed due to technological failures or incomplete survey instruments.
- Measurement errors will have different implications depending on
  - whether the measurement error relates to the dependent or independent variables;
  - the form the measurement error takes.

Measurement Error on the Dependent Variable

- Measurement error on the dependent variable is typically not a serious problem.
- Suppose that the true dependent variable of interest is $y*$, with the regression model taking the form:

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_x x_k + u \tag{26}$$

- Let y denote the observed value for y^*, with

$$e_0 = y - y^* \tag{27}$$

  denoting the measurement error.
- Using the fact that $y^* = y - e_0$, we have from our population regression model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_x x_k + (e_0 + u) \tag{28}$$

Measurement Error on the Dependent Variable (cont'd)

- The key to unbiasedness and consistency of the OLS is that this new composite error term $\tilde{u} = e_0 + u$ have a zero (or constant) conditional mean; i.e.,

$$a = E(\tilde{u}|\boldsymbol{x}) = E(e_0|\boldsymbol{x}) + E(u|\boldsymbol{x}) = E(e_0|\boldsymbol{x}). \tag{29}$$

  where a is a constant.

  Typically, a is assumed to be zero. If it is not, the intercept will be biased, but not the slopes.
- This assumption says that the measurement error is not linked to any of the regressors.
- This could be a problem in some settings (e.g., TOU meters), but is typically not a strong assumption.
- One effect of the measurement error is to increase the error variance, with $Var(\tilde{u}) = Var(e_0 + u) = Var(e_0) + Var(u) > Var(u)$ under the usual assumption that e_0 is uncorrelated with u.

Measurement Error in an Explanatory Variable

- Measurement errors in regressors is typically a more serious problem.
- Consider the single regressor case, with $y = \beta_0 + \beta_1 x_1^* + u$, where x_1^* is the true value of the regressor.
- The observed counterpart is given by x_1, with measurement error

$$e_1 = x_1 - x_1^* \tag{30}$$

- Using the fact that $x_1^* = x_1 - e_1$, we can then rewrite our population model as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1^* + u \\ &= \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1) \\ &= \beta_0 + \beta_1 x_1 + \tilde{u} \end{aligned} \tag{31}$$

  where $\tilde{u} = u - \beta_1 e_1$.
- Assuming $E(e_1) = 0$, the questions is whether $E(e_1|\boldsymbol{x})$=0.

Assumption 1

- Two assumptions are typically made regarding the nature of the measurement error.
- Assumption 1: $Cov(x_1, e_1) = 0$ (or the stronger version $E(e_1|x_1) = 0$).

  This assumes that the observed regressor is uncorrelated with the measurement error.
- With this assumption, we get $E(\tilde{u}|\boldsymbol{x}) = E(u|\boldsymbol{x}) - \beta_1 E(e_1|\boldsymbol{x}) = 0$;

  ...i.e., the model with x_1 still satisfies the zero conditional mean assumption, so that the properties of the OLS estimator remain intact.
- As with measurement error in the dependent variable, however, we do end up with a larger error variance.

Assumption 2 - The Classical Errors-in-Variable (CEV) Model

- The more common assumption in terms of measurement errors is that the measurement error is uncorrelated with the true variable; i.e.
$$Cov(x_1^*, e_1) = 0 \tag{32}$$
or the stronger version: $E(e_1|x_1^*) = 0$.
- This is know as **classical errors-in-variables**.
- This is a natural assumption if we have $x_1 = x_1^* + e_1$ and the two terms on the right-hand side are uncorrelated.
- The problem under these conditions is that x_1 and e_1 are now *necessarily* correlated, since
$$Cov(x_1, e_1) = E(x_1 e_1) = E[(x_1^* + e_1)e_1] = E(e_1^2) = \sigma_{e_1}^2. \tag{33}$$
- This in turn implies that
$$Cov(x_1, \tilde{u}) = E[x_1(u - \beta_1 e_1)] = E(x_1 u) - \beta_1 E(x_1 e_1) = -\beta_1 \sigma_{e_1}^2 \tag{34}$$

The Impact of CEV

- Because the zero conditional mean assumption does not hold under CEV, OLS will be a biased and inconsistent estimator in this case.
- In particular, one can show in the single regressor context that
$$plim(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_{e_1}^2} \right). \tag{35}$$
- This, in turn, implies that
$$|plim(\hat{\beta}_1)| < |\beta_1|. \tag{36}$$
i.e., $\hat{\beta}$ is always closer to zero than β (in large samples).
- This is referred to as **attenuation bias**.
- The following example looks at a regression of birth weight on family income with a CEV error $e_1 \sim \mathcal{N}(0, \sigma_{e_1}^2)$

Measurement Error Effects

	no err	$\sigma_{e_1} = 10$	$\sigma_{e_1} = 20$	$\sigma_{e_1} = 40$	$\sigma_{e_1} = 80$
faminc	0.118				
	(0.029)**				
faminc10		0.100			
		(0.026)**			
faminc20			0.043		
			(0.020)*		
faminc40				0.018	
				(0.013)	
faminc80					0.003
					(0.007)
_cons	115.265	115.776	117.441	118.192	118.589
	(1.002)**	(0.938)**	(0.809)**	(0.654)**	(0.592)**
R^2	0.01	0.01	0.00	0.00	0.00
N	1,388	1,388	1,388	1,388	1,388

* $p < 0.05$; ** $p < 0.01$

Additional Notes on Measurement Error

- The above results for measurement error under assumption 1 generalize readily to multiple regressors.
- Under assumption 2 (CEV), OLS will continue to be biased and inconsistent.
  - There will continue to be attenuation bias for the mis-measured variable's parameter.
  - The impact on other parameters is generally case dependent.

Missing Data

- There will often be cases in which there are individual variables missing for a subset of individuals.
- In the context of survey data, this is referred to as "item nonresponse."
- If the nonresponse is random in nature, then the subset of observations with complete data will still represent a random sample and OLS will be unbiased and consistent.

  . . . The only real consequence is that we will have a smaller sample.
- There are a variety of procedures to "fill-in" the missing data, including
  - hotdecking;
  - multiple imputation;
- If the nonresponse is systematic in some fashion, then the trimmed dataset is a **nonrandom sample** from the population.

Nonrandom Samples

- The extent to which a nonrandom sample is problematic depends on how it is nonrandom.
- If the sample has been chosen on the basis of the independent (exogenous) variables of the model, then it falls into the category of **exogenous sample selection**.
  - This might arise if sampling is done, for example, by income or age groups.
  - Wooldridge gives the example:

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u \tag{37}$$

  - It turns out that OLS will still be unbiased, as long as we including the stratification variables as regressors in the model.

Endogenous Sample Selection

- A bigger problem arises if the sample selection is based on the dependent variables.
- Examples of endogenous sample selection include:
  - On-site sampling (intercepting visitors to a recreation site);
  - Sample truncation (e.g., in a model of wealth, collecting data only on those below the poverty level);
  - Survey nonresponse.
- Dealing with endogenous sampling is, in general, more challenging and will be covered somewhat later in the course.

Outliers and Influential Observations

- It is always good practice to start by simply summarizing your data so as to catch coding errors or outliers that might be problematic.
- It can also be useful to examine the residuals from a regression to see if any pattern emerges or if an outliers emerge, though it is not always clear what to do with such observations.