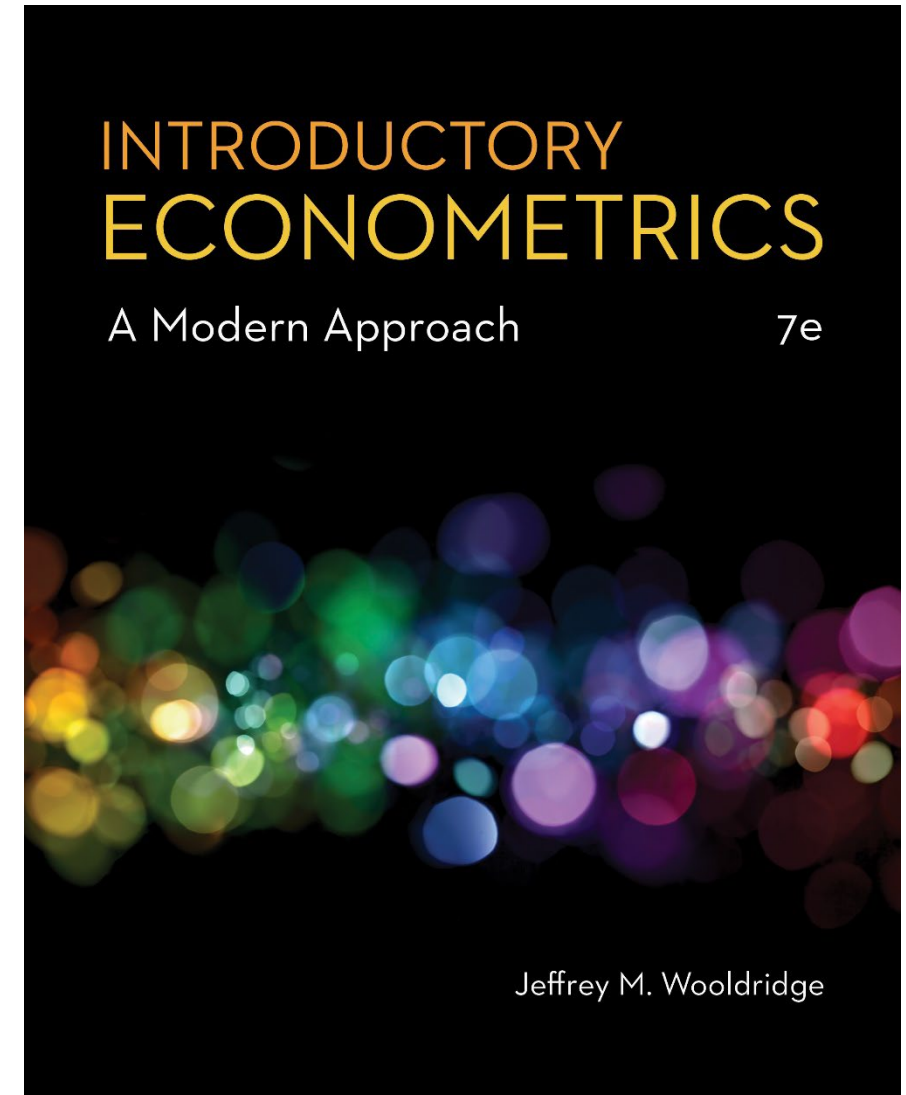


Chapter 9

More on Specification and Data Issues



More on Specification and Data Issues (1 of 20)

- **Tests for functional form misspecification**
 - One can always test whether explanatory should appear as squares or higher order terms by testing whether such terms can be excluded.
 - Otherwise, one can use general specification tests such as RESET
- Regression specification error test (**RESET**)
 - The idea of RESET is to include squares and possibly higher order fitted values in the regression (similarly to the reduced White test).

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

- Test for the exclusion of the \hat{y} terms. If they cannot be excluded, this is evidence for omitted higher order terms and interactions, i.e. for misspecification of functional form.

More on Specification and Data Issues (2 of 20)

- **Example: Housing price equation**

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

$$\Rightarrow F_{2,(88-3-2-1)} = 4.67, p\text{-value} = .012 \leftarrow \text{Evidence for misspecification}$$

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

$$\Rightarrow F_{2,(88-3-1-2)} = 2.56, p\text{-value} = .084 \leftarrow \text{Less evidence for misspecification}$$

- **Discussion**

- One may also include higher order terms, which implies complicated interactions and higher order terms of all explanatory variables.
- RESET provides little guidance as to where misspecification comes from.


More on Specification and Data Issues (3 of 20)

- **Testing against nonnested alternatives**

Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

Model 2: $y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u$

Which specification
is more appropriate?



Define a general model that contains both models as subcases and test:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \log(x_1) + \beta_4 \log(x_2) + u$$

- **Discussion**

- Can always be done; however, a **clear** winner need not emerge
- Cannot be used if the models differ in their definition of the dependent variable

More on Specification and Data Issues (4 of 20)

- Using **proxy variables** for unobserved explanatory variables
- Example: Omitted ability in a wage equation

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

In general, the estimates for the returns to education and experience will be biased because one has omitted the unobservable ability variable. Idea: find a proxy variable for ability which is able to control for ability differences between individuals so that the coefficients of the other variables will not be biased. A possible proxy for ability is the IQ score or similar test scores.

- General approach to use proxy variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3$$

Omitted variable, e.g. ability

Regression of the omitted variable on its proxy

More on Specification and Data Issues (5 of 20)

- **Assumptions necessary for the proxy variable method to work**

- The proxy is “just a proxy” for the omitted variable, it does not belong into the population regression, i.e. **it is uncorrelated with its error.**

$$\text{Corr}(x_3, u) = 0$$

← If the error and the proxy were correlated, the proxy would actually have to be included in the population regression function

- The proxy variable is a “good” proxy for the omitted variable, i.e. using other variables in addition will not help to predict the omitted variable.

$$E(x_3^* | x_1, x_2, x_3) = E(x_3^* | x_3) = \delta_0 + \delta_3 x_3$$

$$\Rightarrow \text{Corr}(x_1, v_3) = \text{Corr}(x_2, v_3) = 0$$

← Otherwise x_1 and x_2 would have to be included in the regression for the omitted variable

More on Specification and Data Issues (6 of 20)

- **Under these assumptions, the proxy variable method works:**

$$\Rightarrow y = (\beta_0 + \beta_3 \delta_0) + \beta_1 x_1 + \beta_2 x_2 + (\beta_3 \delta_3) x_3 + (u + \beta_3 v_3)$$

In this regression model, the error term is uncorrelated with all the explanatory variables. As a consequence, all coefficients will be correctly estimated using OLS. The coefficients for the explanatory variables x_1 and x_2 will be correctly identified. The coefficient for the proxy variable may also be of interest (it is a multiple of the coefficient of the omitted variable).

- Discussion of the proxy assumptions in the wage example
 - Assumption 1: Should be fulfilled as IQ score is not a direct wage determinant; what matters is how able the person proves at work.
 - Assumption 2: Most of the variation in ability should be explainable by variation in IQ score, leaving only a small rest to educ and exper.

More on Specification and Data Issues (7 of 20)

TABLE 9.1 Dependent Variable: $\log(\text{wage})$

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

© Cengage Learning, 2016

- As expected, the measured return to education decreases if IQ is included as a proxy for unobserved ability.
- The coefficient for the proxy suggests that ability differences between individuals are important (e.g. +15 points IQ score are associated with a wage increase of 5.4 percentage points).
- Even if IQ score imperfectly soaks up the variation caused by ability, including it will at least reduce the bias in the measured return to education. Why?
- No significant interaction effect between ability and education.

More on Specification and Data Issues (8 of 20)

- **Using lagged dependent variables as proxy variables**

- In many cases, omitted unobserved factors may be proxied by the value of the dependent variable from an earlier time period.

- Example: City crime rates

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

- Including the past crime rate will at least partly control for the many omitted factors that also determine the crime rate in a given year.
- Another way to interpret this equation is that one compares cities which had the same crime rate last year; this avoids comparing cities that differ very much in unobserved crime factors.

More on Specification and Data Issues (9 of 20)

• Models with random slopes (= random coefficient models)

$$y_i = (\alpha + c_i) + (\beta + d_i)x_i$$

← The model has a **random intercept and a random slope**

Average intercept Random component Average slope Random component

$$\Rightarrow y_i = \alpha + \beta x_i + (c_i + d_i x_i)$$

← Error term

← The individual random components are independent of the explanatory variable

Assumptions: $E(c_i|x_i) = E(d_i|x_i) = 0$

$$\Rightarrow E(c_i + d_i x_i | x_i) = 0$$

$$\Rightarrow \text{Var}(c_i + d_i x_i | x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$$

← WLS or OLS with robust standard errors will consistently estimate the average intercept and average slope in the population

More on Specification and Data Issues (10 of 20)

- Measurement error in an explanatory variable**

$$x_1 = x_1^* + e_1 \quad \leftarrow \text{Mis-measured value} = \text{True value} + \text{Measurement error}$$

$$y = \beta_0 + \beta_1 x_1^* + \dots + \beta_k x_k + u \quad \leftarrow \text{Population regression}$$

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + (u - \beta_1 e_1) \quad \leftarrow \text{Estimated regression}$$

Classical errors-in-variables assumption: $Cov(x_1^*, e_1) = 0$ \leftarrow Error uncorrelated to true value

$$\Rightarrow Cov(x_1, u - \beta_1 e_1) = -\beta_1 Cov(x_1, e_1) = -\beta_1 \sigma_{e_1}^2 \quad \leftarrow \text{The mismeasured variable } x_1 \text{ is correlated with the error term}$$

More on Specification and Data Issues (11 of 20)

- **Consequences of measurement error in an explanatory variable**

- Under the classical errors-in-variables assumption, OLS is biased and inconsistent because the mismeasured variable is endogenous.
- One can show that the inconsistency is of the following form:

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

← This factor (which involves the error variance of a regression of the true value of x_1 on the other explanatory variables) will always be between zero and one

- The effect of the mismeasured variable suffers from **attenuation bias**, i.e. the magnitude of the effect will be attenuated towards zero.
- In addition, the effects of the other explanatory variables will be biased

More on Specification and Data Issues (12 of 20)

- **Missing data and nonrandom samples**
- Missing data as sample selection
 - Missing data is a special case of **sample selection** (= nonrandom sampling) as the observations with missing information cannot be used.
 - If the sample selection is based on independent variables there is no problem as a regression conditions on the independent variables.
 - In general, sample selection is no problem if it is uncorrelated with the error term of a regression (= **exogenous sample selection**).
 - Sample selection is a problem, if it is based on the dependent variable or on the error term (= **endogenous sample selection**).

More on Specification and Data Issues (13 of 20)

- The **Missing Indicator Method (MIM)**
- Suppose we are missing some information on an explanatory variable

Full information: $y, x_1, \dots, x_{k-1};$

Missing some information: x_k

- MIM creates two new variables:

$z_{ik} = x_{ik}$ whenever x_{ik} is observed and $z_{ik} = 0$ when x_{ik} is missing

m_{ik} is a missing data indicator equal to 1 if x_{ik} is missing and 0 otherwise.

Regress y_i on $x_{i1}, \dots, x_{ik-1}, z_{ik}, m_{ik}$ using all observations

More on Specification and Data Issues (14 of 20)

- **Limitations of the Missing Indicator Method**
- If the missing data mechanism is completely at random (**missing completely at random, MCAR**), then MIM estimator is unbiased and consistent.
- It is a strong assumption
- In this situation, we can run the regression using non-missing sample (same treatment in MAR).

Omitting m_{ik} is the same as assuming $x_{ik} = 0$ whenever it is missing.

More on Specification and Data Issues (15 of 20)

- **Example for exogenous sample selection**

$$savings = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u$$

If the sample was nonrandom in the way that certain age groups, income groups, or household sizes were over- or undersampled, this is not a problem for the regression because it examines the savings for subgroups defined by income, age, and hh-size. The distribution of subgroups does not matter.

- **Example for endogenous sample selection**

$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u$$

If the sample is nonrandom in the way individuals refuse to take part in the sample survey if their wealth is particularly high or low, this will bias the regression results because these individuals may be systematically different from those who do not refuse to take part in the sample survey.

More on Specification and Data Issues (16 of 20)

- Interpreting the meaning of **Missing at Random (MAR)**
- The term Missing at Random (MAR) means that the missingness is unrelated to the unobserved error u , but it is related to the explanatory variables (x_1, \dots, x_k) .
- Missing Completely at Random (MCAR) means that missingness is unrelated to both u and x_1, \dots, x_k .

More on Specification and Data Issues (17 of 20)

- **Outliers and influential observations**

- Extreme values and outliers may be a particular problem for OLS because the method is based on squaring deviations.
- If outliers are the result of mistakes that occurred when keying in the data, one should just **discard the affected observations**.
- If outliers are the result of the data generating process, the decision whether to discard the outliers is not so easy.

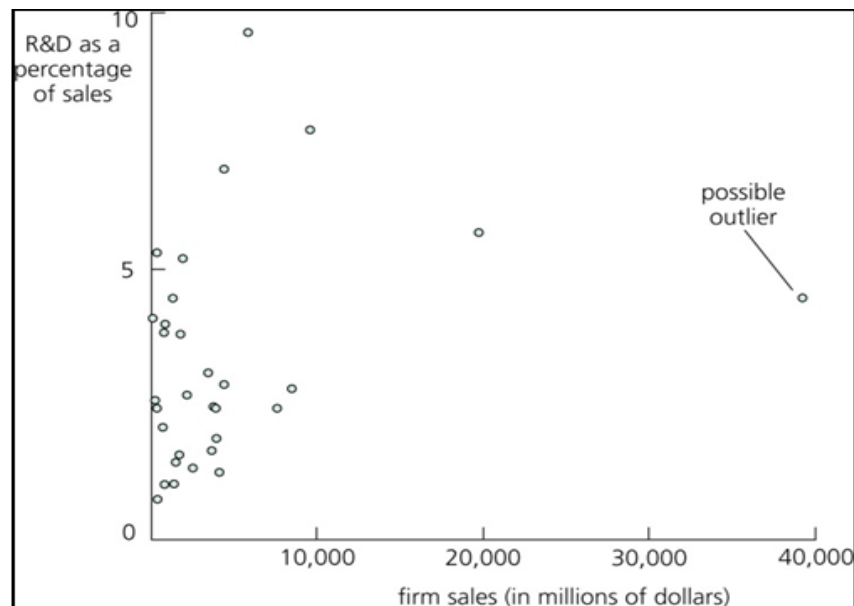
- Example: R&D intensity and firm size

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + u$$

More on Specification and Data Issues (18 of 20)

• Example: R&D intensity and firm size (cont.)

rdchem.dta



$$\widehat{rdintens} = 2.63 + .00005 \text{ sales} + .045 \text{ profmarg}$$

$$(0.59) \quad (.00004) \quad (.046)$$

$$n = 32, R^2 = .0761, \bar{R}^2 = .0124$$

The regression without the outlier makes more sense.

$$\widehat{rdintens} = 2.30 + .00019 \text{ sales} + .048 \text{ profmarg}$$

$$(0.59) \quad (.00008) \quad (.045)$$

$$n = 31, R^2 = .1728, \bar{R}^2 = .1137$$

The outlier is not the result of a mistake: One of the sampled firms is much larger than the others.

More on Specification and Data Issues (19 of 20)

- **Least absolute deviations estimation (LAD)**

- The least absolute deviations estimator minimizes the sum of absolute deviations (instead of the sum of squared deviations, i.e. OLS)

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

- It may be more robust to outliers as deviations are not squared.
- The least absolute deviations estimator estimates the parameters of the conditional median (instead of the conditional mean with OLS).
- The least absolute deviations estimator is a special case of quantile regression, which estimates parameters of conditional quantiles.