FINM3123 Introduction to Econometrics

Chapter 9

More on Specification and Data Issues

Multiple Regression Analysis: Specification and Data Issues

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)

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

Test for the exclusion of these terms. If they cannot be excluded, this is evidence for omitted higher order terms and interactions, i.e. for misspecification of functional form.

Multiple Regression Analysis: Specification and Data Issues

Example: Housing price equation

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

$$\Rightarrow \quad F_{2,(88-3-2-1)} = 4.67, p - value = .012$$

Evidence for misspecification

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

$$\Rightarrow \quad F_{2,(88-3-1-2)} = 2.56, p - value = .084$$

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

Multiple Regression Analysis: Specification and Data Issues

Testing against non-nested alternatives

Which specification is more appropriate?

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

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

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

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

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 variables

Multiple Regression Analysis: Specification and Data Issues

<u>Using proxy variables for unobserved explanatory variables</u>

Example: Omitted ability in a wage equation

Replace by proxy

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

In general, the estimates for the returns to education and experience will be biased because one has omitted the unobservable *ability* variable. <u>Idea</u>: 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 using proxy variables

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

Omitted variable, e.g. ability

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

Regression of the omitted variable on its proxy

Multiple Regression Analysis: Specification and Data Issues

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

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

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 additional variables will not help to predict the omitted variable

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

$$\Rightarrow \quad Corr(x_1, v_3) = Corr(x_2, v_3) = 0$$

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

Multiple Regression Analysis: Specification and Data Issues

Under these assumptions, the proxy variable method works:

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

In this regression model, the error term is uncorrelated with all 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 is 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*

Multiple Regression Analysis: Specification and Data Issues

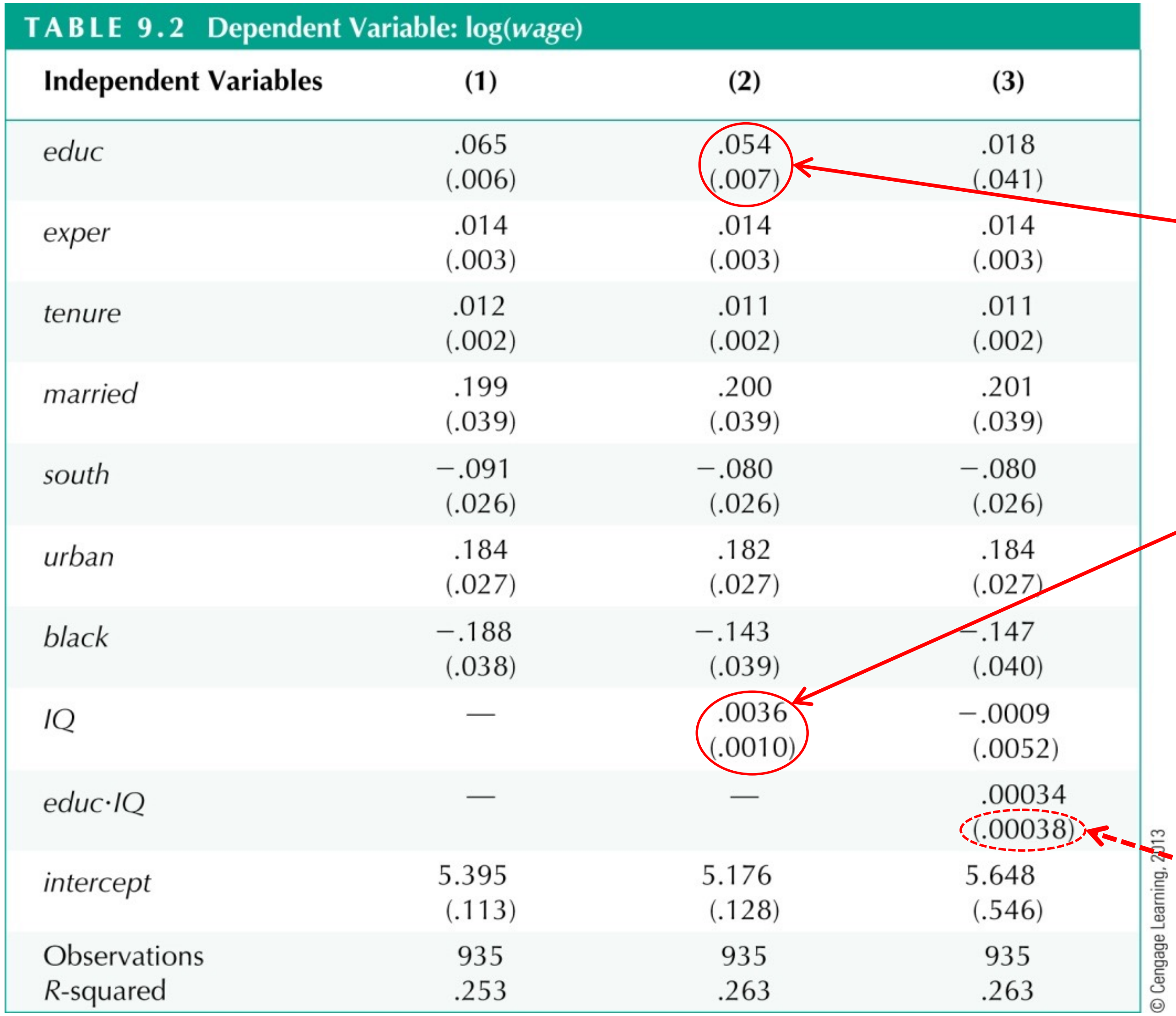


TABLE 9.2 Dependent Variable: log(*wage*)

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

© Cengage Learning, 2013

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.

No significant interaction effect between ability and education.

Multiple Regression Analysis: Specification and Data Issues

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

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

- 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

Multiple Regression Analysis: Specification and Data Issues

Models with random slopes (= random coefficient models)

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

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

Average intercept

Random component

Average slope

Random component

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

Error term

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

The individual random components are independent of the explanatory variable

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

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

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

Multiple Regression Analysis: Specification and Data Issues

<u>Properties of OLS under measurement error</u>

Measurement error in the <u>dependent</u> variable

$$y = y^* + \enclose{circle}{e_0}$$ Mismeasured value = true value + measurement error

$$\boxed{y^*} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$ Population regression

$$\Rightarrow \boxed{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \boxed{(u + e_0)}$$ Estimated regression

Consequences of measurement error in the dependent variable

- Estimates will be less precise because the error variance is higher
- Otherwise, OLS will be unbiased and consistent (as long as the measurement error is unrelated to the values of the explanatory variables)

Multiple Regression Analysis: Specification and Data Issues

Measurement error in an <u>explanatory variable</u>

$$x_1 = x_1^* + \boxed{e_1}$$

Mismeasured value = True value + Measurement error

$$y = \beta_0 + \beta_1 \boxed{x_1^*} + \cdots + \beta_k x_k + u$$

Population regression

$$\Rightarrow \quad y = \beta_0 + \beta_1 \boxed{x_1} + \cdots + \beta_k x_k + \boxed{(u - \beta_1 e_1)}$$

Estimated regression

<u>**Classical errors-in-variables (CEV) assumption**</u>: $\boxed{Cov(x_1^*, e_1) = 0}$

Error unrelated to true value

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

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

<u>The mismeasured variable x_1 is correlated with the error term!</u>

Multiple Regression Analysis: Specification and Data Issues

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:

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

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

Multiple Regression Analysis: Specification and Data Issues

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 the independent variables there is no problem as a regression conditions on the independent variables
- In general, sample selection is not a problem if it is uncorrelated with the error term of the 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)

Multiple Regression Analysis: Specification and Data Issues

Example of exogenous sample selection

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

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 the savings model is the same for all subgroups defined by income, age, and household size. The distribution of subgroups does not matter.

Example of 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.

Multiple Regression Analysis: Specification and Data Issues

Outliers and influential observations

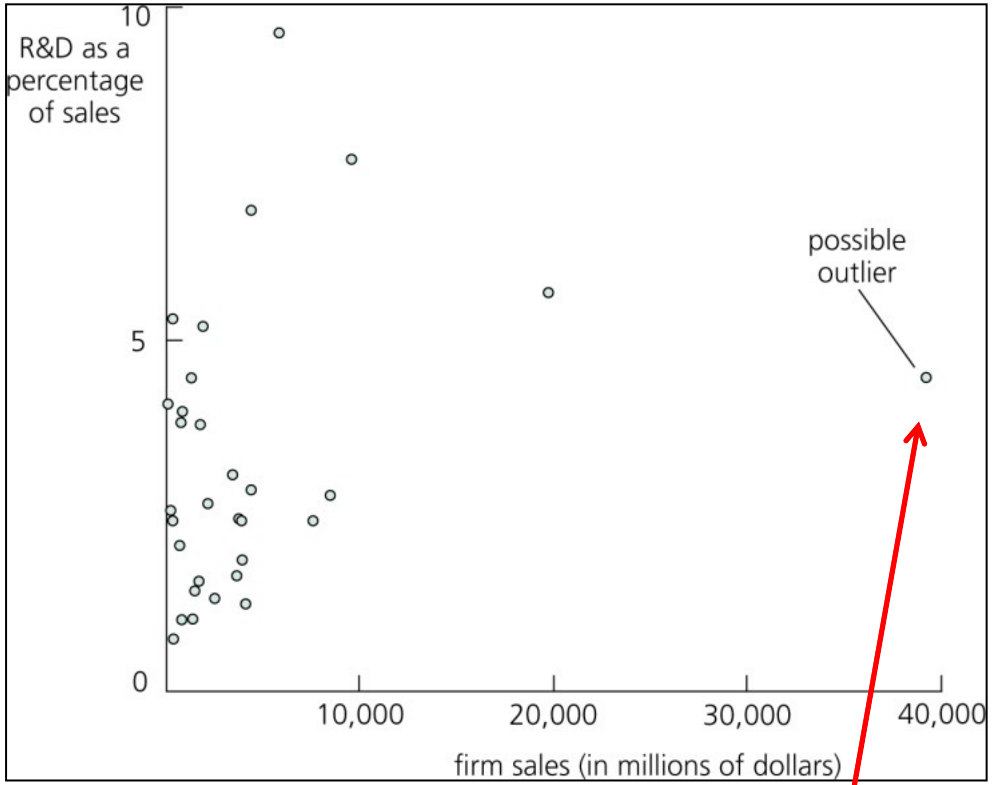
- Extreme values and outliers may be a particular problem for OLS because the method is based on squaring deviations
- If outliers are the result of mistakes that occurred when keying in the data, one should just discard the affected observations
- If outliers are the result of the data generating process,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

Example: R&D intensity and firm size

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

Multiple Regression Analysis: Specification and Data Issues

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



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

$$\widehat{rdintens} = \underset{(0.59)}{2.63} + \underset{(.00004)}{.00005}\ sales + \underset{(.046)}{.045}\ profmarg$$

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

$$\widehat{rdintens} = \underset{(0.59)}{2.30} + \underset{(.00008)}{.00019}\ sales + \underset{(.045)}{.048}\ profmarg$$

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

The regression without the outlier makes more sense.

Multiple Regression Analysis: Specification and Data Issues

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)

$$\min \sum_{i=1}^{n} |y_i - b_0 - b_1 x_1 - \cdots - b_k x_k|$$

- It is 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

Summary

- Tests for functional form misspecification
  - RESET
  - Test against nonnested alternatives
    - Two approaches
    - A clear winner need not emerge
    - Dependent variables must be the same
- Proxy variables
  - Two assumptions

Summary

- Measurement error
  - Measurement error in the dependent variable
    - Estimates will be less precise
    - OLS estimators can still be unbiased and consistent (as long as the measurement error is unrelated to the values of explanatory variables)
  - Measurement error in an independent variable
    - CEV assumption
    - Under CEV assumption, OLS is biased and inconsistent
    - Attenuation bias: $plim(\hat{\beta}_1)$ is always closer to zero than β_1 is.

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

Summary

- Missing data and nonrandom samples
    - Exogenous sample selection
        - Can still get unbiased and consistent estimators
    - Endogenous sample selection
        - Biased and inconsistent estimators
- Outliers
    - In cases where one or several data points substantially change the results, OLS results should probably be reported with and without outliers.
- Least absolute deviation estimation (LAD)
    - Minimizes the sum of absolute deviations
    - Robust to outliers
    - Estimates the parameter of the conditional median
    - LAD estimates are not available in closed form