Ch09. More on Specification and Data Issues

Ping Yu

HKU Business School
The University of Hong Kong

Functional Form Misspecification

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 regression specification error test (RESET) of Ramsey [photo here].

- Ramsey, J.B., 1969, Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis, *Journal of the Royal Statistical Society*, *Series B*, 31, 350-371.
- Ramsey, J.B., 1970, Models, Specification Error, and Inference: A Discussion of Some Problems in Econometric Methodology, *Bulletin of the Oxford Institute of Economics and Statistics*, 32, 301-318.

James B. Ramsey (1937-), NYU, 1968UW-MadisonPhD

a: RESET as a General Test for Functional Form Misspecification

- Let $\widehat{y}$ be the fitted value from the initial linear regression.
- Basic Idea: include squares and possibly higher order fitted values in the regression (similarly to the reduced White test),

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

  and then test whether $\delta_1 = \delta_2 = 0$ by the F test.
- If $\widehat{y}^2$ and $\widehat{y}^3$ cannot be excluded, this is evidence for omitted higher order terms and interactions, i.e., for misspecification of functional form.

- 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.

Example: Housing Price Equation

- We estimate two models for housing prices:

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

  in level form and

$$lprice = \beta_0 + \beta_1 llotsize + \beta_2 lsqrft + \beta_3 bdrms + u$$

  in log form (except *bdrms*).

- For the first specification,

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

  This is evidence of functional form misspecification.

- For the second specification,

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

  This is less evidence for misspecification.

b: Tests against Nonnested Alternatives

Mizon-Richard Test

- Suppose we want to know which of the following two specifications is more appropriate:
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{1}$$
or
$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u. \tag{2}$$
- Method I: Mizon, G.E. and J.F. Richard, 1986, The Encompassing Principle and Its Application to Testing Nonnested Hypotheses, *Econometrica*, 54, 657-678. [photo here]
- Basic Idea: construct a comprehensive model that contains each model as a special case and then to test the restrictions that led to each of the models.
- Specifically, let
$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 \log(x_2) + u.$$
- Test $H_0 : \gamma_3 = \gamma_4 = 0$ for model (1) and test $H_0 : \gamma_1 = \gamma_2 = 0$ for model (2).

Grayham E. Mizon (?-),
Southampton, 1972LSEPhD

Jean-Francois Richard (1943-),
Pittsburgh, 1973LouvainPhD

continue

Davidson-MacKinnon test

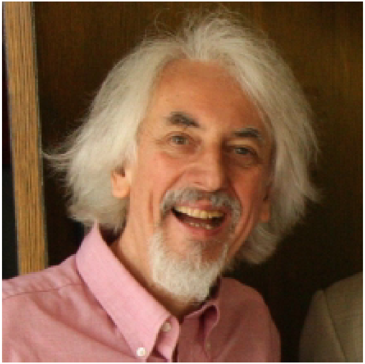
- Method II: Davidson, R. and J.G. MacKinnon, 1981, Several Tests of Model Specification in the Presence of Alternative Hypotheses, *Econometrica*, 49, 781-793. [photo here]
- Basic Idea: if (1) is true, then the fitted values from the other model, (2), should be insignificant in (1), and vise versa.
- Specifically, run the regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \widehat{\widehat{y}} + error,$$

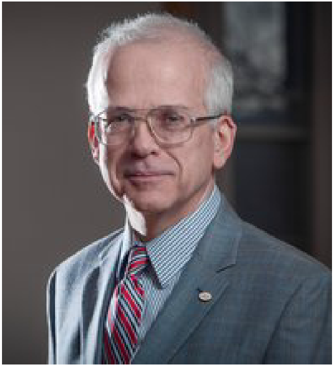
  and test $H_0 : \theta_1 = 0$ for model (1), where $\widehat{\widehat{y}}$ is the predicted value from model (2).
- Similarly, run the regression

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_2 \widehat{y} + error,$$

  and test $H_0 : \theta_2 = 0$ for model (2), where $\widehat{y}$ is the predicted value from model (1).



Russell Davidson (1941-),
McGill, 1977UBCPhD



James G. MacKinnon (1951-),
Queen, 1975PrincetonPhD

Discussion

- Problem I: In nonnested testing, a clear winner need not emerge. Both models could be rejected or neither model could be rejected.
  - In the latter case, we can use the adjusted R-squared to choose between them. If both models are rejected, more work needs to be done.
- Problem II: Rejecting (1) using, say, the Davidson-MacKinnon test does not mean that (2) is the correct model. Model (1) can be rejected for a variety of functional form misspecifications.
- Problem III: Cannot be used if the models differ in their definition of the dependent variables, e.g., y versus $\log(y)$. [see Chapter 6 for goodness-of-fit measures that can be compared.]

Using Proxy Variables for Unobserved Explanatory Variables

Example: Omitted Ability in a Wage Equation

- Suppose the true wage equation is

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

- 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 Using Proxy Variables

- Suppose
$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3^*+u,$$
where x_3^* is an unobservable omitted variable but has a proxy variable x_3 for it.
- Regressing the omitted variable on its proxy, we have
$$x_3^*=\delta_0+\delta_3x_3+v_3,$$
where δ_3 is usually positive for x_3 to be a suitable proxy for x_3^*.
- Assumption I: The proxy is "just a proxy" for the omitted variable, it does not belong to the population regression, i.e., it is uncorrelated with its error:
$$Cov\left(x_3,u\right)=0.$$
- If the error and the proxy were correlated, the proxy would actually have to be included in the original population regression function.
- Assumption II: 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:
$$E\left[x_3^*|x_1,x_2,x_3\right]=E\left[x_3^*|x_3\right]=\delta_0+\delta_3x_3,$$
which implies $Cov\left(x_1,v_3\right)=Cov\left(x_2,v_3\right)=0$; in other words, all the correlations between x_3^* and $\left(x_1,x_2\right)$ are through x_3.
- Otherwise, x_1 and x_2 would have to be included in the regression for the omitted variable.

Plug-in Solution to the Omitted Variables Problem

- Plug in x_3 for x_3^* and run regression of y on (x_1, x_2, x_3).
- The relationship between y and (x_1, x_2, x_3) is

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (\delta_0 + \delta_3 x_3 + v_3) + u \\ &= (\beta_0 + \beta_3 \delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \delta_3 x_3 + (\beta_3 v_3 + u), \end{aligned}$$

  where $Cov(x_1, v_3) = Cov(x_2, v_3) = Cov(x_3, v_3) = 0$ (why?) and $Cov(x_1, u) = Cov(x_2, u) = Cov(x_3, u) = 0$ (why?).
- In this regression model, the error term $\beta_3 v_3 + u$ 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).

Revisit the Wage Example

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

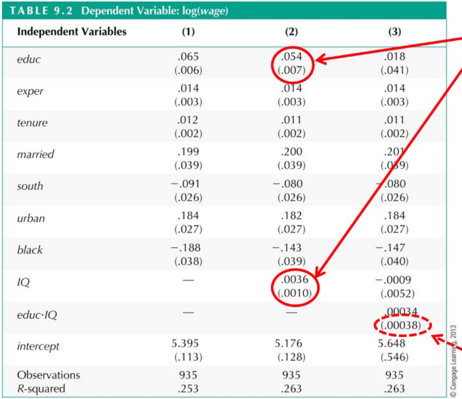


TABLE 9.2 Dependent Variable: log(*wage*)

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

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.

a: 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):

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

  where
  $crime =$ a measure of per capita crime
  $unem =$ the city unemployment rate
  $expend =$ per capita spending on law enforcement
- 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.

Models with Random Slopes (Random Coefficient Models)

History of Random Coefficient Models (RCMs)

P.A.V.B. Swamy (1934-)
Federal Reserve, 1968UW-MadisonPhD

RCM

- In the return-to-schooling example, the return to education may vary across individuals, i.e.,

$$y_i = \beta_0 + b_i x_i + u_i \equiv a_i + b_i x_i.$$

  where $y_i = \log(wage_i)$ and $x_i = edu_i$.

- Generally, write $a_i = \alpha + c_i$ with $\alpha = E[a_i]$ and $b_i = \beta + d_i$ with $\beta = E[b_i]$, where α is the average intercept and c_i is the random component, the average slope β is called average partial effect (APE) or average marginal effect (AME).
- Thus,

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

- To guarantee $E[u_i|x_i] = 0$, assume $E[c_i|x_i] = E[d_i|x_i] = 0$, i.e., the individual random components are mean independent of the explanatory variable.
- Now, $E[c_i + d_i x_i|x_i] = 0$ and $Var(c_i + d_i x_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$, i.e., the model is heteroskedastic, which implies that WLS or OLS with robust standard errors will consistently estimate the average intercept and average slope in the population.

Properties of OLS under Measurement Error

a: Measurement Error in the Dependent Variable

- Suppose $y = y^* + e_0$, i.e., mismeasured value = true value + measurement error.
- If the population regression is

  $$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u,$$

  satisfying the Gauss-Markov assumptions, then the estimated regression is

  $$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u + e_0.$$

- For simplicity, assume e_0 is independent of **x** and u with mean zero.
- Consequence: Estimates will be less precise because the error variance is higher ($Var\left(u + e_0\right) = Var\left(u\right) + Var\left(e_0\right) > Var\left(u\right)$); otherwise, OLS will be unbiased and consistent (as long as the measurement error is unrelated to the values of the explanatory variables, $E\left[e_0|\mathbf{x}\right] = E\left[e_0\right] = 0$).

b: Measurement Error in an Explanatory Variable

- Suppose $x_1 = x_1^* + e_1$, i.e., mismeasured value = true value + measurement error.
- If the population regression is

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

then the estimated regression is

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1).$$

- Classical error-in-variables assumption (CEV): $Cov(x_1^*, e_1) = 0$, i.e., the measurement error is uncorrelated with the true value.
- Now,

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

and

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

i.e., the mismeasured variable x_1 is correlated with the error term!

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

$$\begin{aligned} E\left[\widehat{\beta}_1\right] &= \beta_1 + \frac{Cov\left(x_1, u-\beta_1 e_1\right)}{Var\left(x_1\right)} = \beta_1 - \beta_1 \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\ &= \beta_1\left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right) = \beta_1 \frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}. \end{aligned}$$

- Since the reliability coefficient $\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2+\sigma_{e_1}^2}$ is less than 1, $E\left[\widehat{\beta}_1\right]$ is always closer to zero than β_1, which is called the attenuation bias.
- Besides the magnitude of the effect of the mismeasured variable will be attenuated towards zero, the effects of the other explanatory variables (if they are present and precisely measured) will be biased.
- Hausman (2000): "At MIT I have called this the “Iron Law of Econometrics” – the magnitude of the estimate is usually smaller than expected." [photo here]

History of Measurement Error

Jerry A. Hausman (1946-), MIT, 1973OxfordPhD

Missing Data, Nonrandom Samples, and Outlying Observations

b: 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 **x**, there is no problem as a regression conditions on the independent variables ($E[u|\mathbf{x}] = 0$ still holds for observed **x** values.)
- 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).

Examples of Sample Selection

- Example for exogenous sample selection: Suppose the model of interest is

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

- 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 it examines the savings for subgroups defined by income, age, and household size. The distribution of subgroups does not matter.
- Example for endogenous sample selection: Suppose the model of interest is

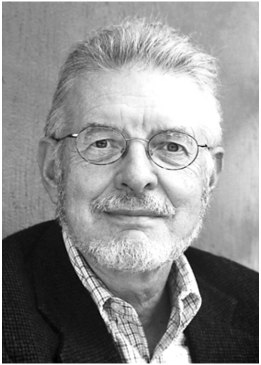
$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u.$$

- 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.

History of Sample Selection

- Heckman, J., 1979, Sample Selection as a Specification Error, *Econometrica*, 47, 153-161.
  - It is one of the most cited papers in econometrics.

James J. Heckman (1944-),
Chicago, 2000NP, 1971PrincetonPhD



Daniel L. McFadden (1937-),
Berkeley, co-winner, 1962MinnesotaPhD

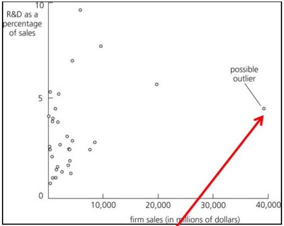
c: Outliers and Influential Observations

- 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 occured 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): The model of interest is

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

where
$rdintens =$ R&D expenditures as a percentage of sales (in millions)
$profmarg =$ profits as a percentage of sales

Example Continue



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

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

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

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

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

The regression without the outlier makes more sense.

Least Absolute Deviations Estimation

LAD

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

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

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

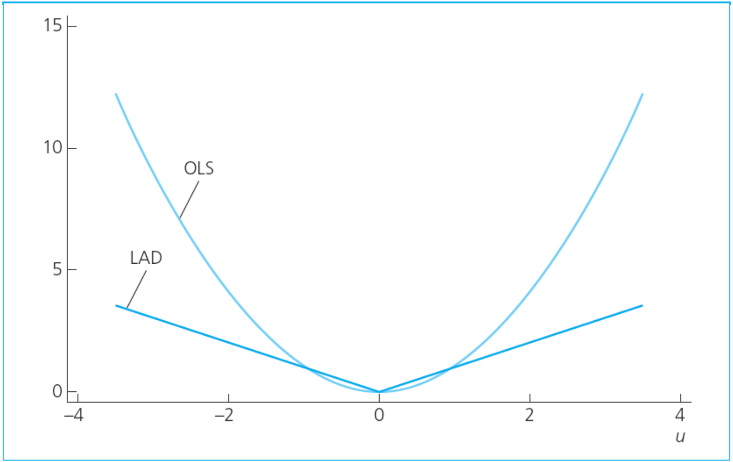
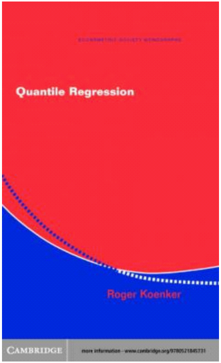


Figure: The OLS and LAD Objective Functions

History of Quantile Regression

- The path-breaking paper: Koenker, R. And G. Bassett, 1978, Regression Quantiles, *Econometrica*, 46, 33-50.


ECONOMETRIC SOCIETY MONOGRAPHS
Quantile Regression
Roger Koenker
CAMBRIDGE
more information - www.cambridge.org/9780521845731

Roger W. Koenker (1947-),
UIUC, 1974MichiganPhD