MISSPECIFICATION and DATA PROBLEMS

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

Econometrics I

Model Specification and Data Problems

- In the previous class we analyzed one failure of Gauss-Markov assumptions: MLR.5 Constant Variance
- Heteroscedasticity does not cause bias or inconsistency in the OLS estimators but causes inefficiency. We learned that it is relatively easy to adjust standard errors and test statistics.
- Now we want to analyze a more serious problem, namely, violation of the assumption of exogeneity (MLR.4). We will examine the case where "the error term u is correlated with one or more of the explanatory variables" (ie, endogeneity).
- Recall that if the x variable is correlated with the error term it is called an **endogenous variable**.
- Recall that when a relevant variable is omitted from the model OLS estimators are biased and inconsistent.
- In the special case that the omitted variable is a function of an explanatory variable in the model, the model suffers from functional form misspecification.

Model Specification and Data Problems

- In this chapter, we will first discuss **functional form misspecification** and how to test for it.
- Then, we will discuss how to use **proxy variables** to mitigate omitted variable bias.
- We will also discuss problems caused by **measurement errors** in dependent and explanatory variables.
- We will discuss the problems caused by endogenous variables within the context of OLS estimators. In most cases, endogeneity problem cannot be solved within the OLS framework. We will need consistent estimation methods such as Instrumental Variables and Two Stage Least Squares (2SLS).

Functional Form Misspecification

- A multiple regression model suffers from functional form misspecification when it does not properly account for the relationship between the dependent and observed explanatory variables.
- For example, if we fit a level-level model instead of a log-log model (which is the true specification); or if we omit a quadratic term where we should have added, then the model suffers from functional form misspecification. This, of course, leads to biased and inconsistent $\hat{\beta}_j$.
- Another example: suppose that the return to an additional year of education changes with the gender implying that the model should contain an interaction term. If we omit, for some reason, this interaction term given by $female \times educ$ then the functional form will be misspecified.

Functional Form Misspecification

- How to detect misspecified functional form? We can always use the F-test for the joint exclusion restrictions such joint significance of quadratic terms, interaction terms, etc.
- We can use usual statistical inference procedures to mitigate the functional form misspecification problem.
- Significant quadratic terms may be symptomatic of other functional form problems such as using level of a variable when the log is more appropriate.
- In fact, using log transformation, where appropriate, may work well in practice.

A Test for Functional Form Misspecification

- Is there a general test that can detect functional form misspecification?
- Yes, there are many misspecification tests. We will only examine one of them.
- We will learn "Regression Specification Error Test" or RESET test of Ramsey (1969).
- Ramsey's RESET test is designed to detect if there are any neglected nonlinearities in the model.

Ramsey's RESET Test

- Suppose that in the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

  the assumption MLR.4 (exogenous xs) is satisfied.
- This implies that no nonlinear functions of the independent variables (such as squares and cubes of x_js should be significant when added to the model.
- But, as in the White heteroscedasticity test, adding squares, cubes and cross-products uses up many degrees of freedom. This is a drawback.
- Instead of this, we can add squares and cubes of the fitted values, $\hat{y}^2$, $\hat{y}^3$, into the model and test for the joint significance of added terms using F or LM test.

RESET Test

- The auxiliary regression for the RESET test statistic can be written as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$$

- The null hypothesis of the RESET test says that the model is correctly specified:

$$H_0 : \delta_1 = 0, \delta_2 = 0$$

- In large samples and under the Gauss-Markov assumptions, the usual F restrictions test follows the $F(2, n-k-3)$ distribution.
- If the F statistic is greater than the critical value at a given significance level then we reject the null hypothesis of correct specification. This indicates that there is a functional form misspecification.
- We can also use LM test statistic. The LM test statistic follows the χ^2_2 distribution.

RESET Test Example: House prices, hprice1.gdt

- Level-level model:

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

- Level-level estimation results:

$$\widehat{\text{price}} = \underset{(29.475)}{-21.77} + \underset{(0.0006)}{0.002}\ \text{lotsize} + \underset{(0.013)}{0.123}\ \text{sqrft} + \underset{(9.010)}{13.85}\ \text{bdrms}$$

$$n = 88 \quad R^2 = 0.672$$

- We form our test regression by adding squares and cubes of $\hat{y}$ into the model above.

RESET Test Example: Level-level Model

```
Auxiliary regression for RESET specification test
OLS, using observations 1-88
Dependent variable: price

              coefficient     std. error     t-ratio   p-value
  ------------------------------------------------------------
  const       166.097         317.433         0.5233    0.6022
  lotsize       0.000153723     0.00520304    0.02954   0.9765
  sqrft         0.0175988       0.299251      0.05881   0.9532
  bdrms         2.17490        33.8881        0.06418   0.9490
  yhat^2        0.000353426     0.00709894    0.04979   0.9604
  yhat^3        1.54557e-06     6.55431e-06   0.2358    0.8142

Test statistic: F = 4.668205,
with p-value = P(F(2,82) > 4.66821) = 0.012
```

in GRETL: from the menu within estimation results window:
TESTS-RAMSEY'S RESET-SQUARES and CUBES
RESET Test Result: at 5% significance level we reject the null hypothesis which states that the functional form is correctly specified. Thus, there is functional form misspecification.

RESET Test Example: hprice1.gdt

- Alternative functional form: log-log model (except bdrms)

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

- Log-log estimation results:

$$\widehat{\text{lprice}} = \underset{(0.651)}{-1.297} + \underset{(0.038)}{0.17}\ \text{llotsize} + \underset{(0.093)}{0.70}\ \text{lsqrft} + \underset{(0.028)}{0.037}\ \text{bdrms}$$

$$n = 88 \quad R^2 = 0.643$$

- Now let us calculate RESET test statistic.

RESET Test Example: hprice1.gdt

```
Auxiliary regression for RESET specification test
OLS, using observations 1-88
Dependent variable: lprice

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       87.8849      240.974       0.3647   0.7163
  llotsize    -4.18098      12.5952     -0.3319   0.7408
  lsqrft     -17.3491       52.4899     -0.3305   0.7418
  bdrms       -0.925329      2.76975    -0.3341   0.7392
  yhat^2       3.91024      13.0143      0.3005   0.7646
  yhat^3      -0.192763      0.752080   -0.2563   0.7984

Test statistic: F = 2.565042,
with p-value = P(F(2,82) > 2.56504) = 0.0831
```

RESET Test Result: at 5% significance level, we fail to reject the null hypothesis of correct specification. This indicates that the functional form is correct. We prefer log-log specification.

RESET Test

- A drawback with RESET test is that it provides no real direction on how to proceed if the model is rejected.
- Some have argued that RESET is a very general test for model misspecification, including unobserved omitted variables and heteroscedasticity.
- This conclusion is misguided. If the omitted variable is linearly related to the included variables the RESET test has no power detecting this.
- Also, if the functional form is correct, the RESET test has no power for detecting heteroscedasticity.
- RESET test is just a functional form test. It should not be used for other purposes.

Tests Against Nonnested Alternatives

- There are several tests for functional form misspecification. Consider the following two models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

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

- These are nonnested models. We cannot write one of these model as a special case of the other.
- In this case we cannot use F test.
- As long as the dependent variable is the same, two different approaches have been suggested.
- We can form a bigger model which includes both models as special cases and use F test. This method is suggested by Mizon-Richard.

Tests Against Nonnested Alternatives

- The other method is known as the Davidson-MacKinnon test. This test is based on including the fitted values $\hat{y}$ from one model into the other model as an additional regressor and conducting a t-test.
- We will not examine these tests in detail.
- There are several drawbacks associated with nonnested tests.
- First, these tests may not choose a correct specification. Both models could be rejected or neither model could be rejected.
- If neither model could be rejected, we can use the adjusted R-square to choose between them.
- Second, rejecting one model does not automatically mean that the alternative is correct. The true model may have a completely different specification.
- Third, if the dependent variable is different, for example if one has y and the other has $\log(y)$ as dependent variables, these tests cannot be used. We need to employ more complex testing procedures which we will not discuss here.

Using Proxy Variables for Unobserved Explanatory Variables

- Can we use a proxy variable for an omitted unobserved explanatory variable?
- We know that if the unobserved variable is an important, relevant variable then OLS estimators are biased and inconsistent.
- The question can be rephrased as follows: Can we solve or at least mitigate the omitted variable bias using proxy variables?
- A **Proxy variable** is something that is related to the unobserved variable that we would like to control for.
- Example: recall that in the wage equation we could not observe innate ability. Can we use intelligence quotient (IQ) as a proxy for ability?
- IQ does not have to be the same thing as ability, we know they are not. But what we need is for IQ to be correlated with ability.

Using Proxy Variables

- Consider the following model

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

  y :log(wage), x_1:educ, x_2: exper, x_3^*:ability (unobserved)
- x_3^*: unobserved; x_3: proxy for unobserved variable
- Proxy variable must be related to the unobserved variable, represented by the following simple regression:

$$x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$$

- We need the error term ν_3 because these variables are not exactly related.
- Typically, these variables are positively correlated so that $\delta_3 > 0$.
- If $\delta_3 = 0$ then x_3 cannot be a suitable proxy.

Using Proxy Variables

- How can we use x_3 to get an unbiased or at least consistent estimators?
- We can just pretend that x_3^* and x_3 are the same and run the regression of y on x_1, x_2, x_3. This is called **plug-in solution to the omitted variables problem**.
- How does this approach produce consistent estimators?
- To show this we need to make some assumptions about the error terms u and ν_3.
- The error term, u, is uncorrelated with x_1, x_2 and x_3^*. This is the standard MLR.4 assumption.
- In addition to this, u must be uncorrelated with x_3. Since x_3 is the proxy variable, it is irrelevant in the population model. It is x_3^* that affects y not x_3.

$$\mathsf{E}(u|x_1, x_2, x_3^*, x_3) = \mathsf{E}(u|x_1, x_2, x_3^*) = 0$$

Using Proxy Variables

- The error term ν_3 is uncorrelated with x_1, x_2 and x_3.
- This can be stated as follows:

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

- This says that once x_3 is controlled for the expected value of x_3^* does not depend on x_1 and x_2.
- For example, in the wage equation where IF is the proxy variable for ability this condition becomes

$$\mathsf{E}(ability|educ, exper, IQ) = \mathsf{E}(ability|IQ) = \delta_0 + \delta_3 IQ$$

- This implies that the average level of ability only changes with IQ, not with $educ$ and $exper$. Is this a reasonable assumption?

Using Proxy Variables

- Plugging in $x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$ into the model and rearranging we obtain

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3\nu_3$$

- Let the composite error term be $e = u + \beta_3\nu_3$

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

  where $\alpha_0 = (\beta_0 + \beta_3\delta_0), \alpha_3 = \beta_3\delta_3$
- If the assumptions for the proxy variables are all satisfied then the composite error term e will be uncorrelated with the explanatory variables included in the model. Thus, OLS estimators of $\alpha_0, \beta_1, \beta_2, \alpha_3$ will be consistent.
- The coefficient on IQ, α_3, measures the impact of a one point change in IQ test score on wage.

Using Proxy Variables: Wage2.gdt

- This data set contains information about monthly wages, education, experience, tenure, IQ scores, and several demographic characteristics for a sample of 935 working men in 1980.
- Adding IQ test scores we obtain the following results:

Model 1: OLS, using observations 1–935
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	5.17644	0.128001	40.4407	0.0000
educ	0.0544106	0.00692849	7.8532	0.0000
exper	0.0141458	0.00316510	4.4693	0.0000
tenure	0.0113951	0.00243938	4.6713	0.0000
married	0.199764	0.0388025	5.1482	0.0000
south	−0.0801695	0.0262529	−3.0537	0.0023
urban	0.181946	0.0267929	6.7908	0.0000
black	−0.143125	0.0394925	−3.6241	0.0003
IQ	0.00355910	0.000991808	3.5885	0.0004

Mean dependent var	6.779004	S.D. dependent var	0.421144
Sum squared resid	122.1203	S.E. of regression	0.363152
R^2	0.262809	Adjusted R^2	0.256441

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

Using Lagged Dependent Variables as Proxy Variables

- In some applications (eg, the wage example) we have at least a vague idea about which unobserved factor we want to control.
- In other applications, we suspect that one or more of the independent variables is correlated with an omitted variable, but we have no idea how to obtain a proxy for that omitted variable.
- In such cases, we can include the value of the dependent variable y from an earlier time period, y_{-1}.
- To do this we need the lagged value of the dependent variable. This provides a way of controlling historical factors that cause current differences in dependent variable.
- For example, some cities have had high crime rates in the past Many of the unobserved factors contribute to both high current and past crime rates. Slowly moving components in dependent variable (inertial effects) can be captured by the lagged value.

Using Lagged Dependent Variables as Proxy Variables

- Example: CRIME2.gdt, 1987 crime data for 46 cities, information in 1982 also available
- The model without the lagged crime rate:

$$\widehat{\text{l\_crmrte87}} = \underset{(1.251)}{3.34} - \underset{(0.032)}{0.029}\,\text{unem87} + \underset{(0.173)}{0.203}\,\text{l\_lawexpc87}$$
$$n = 46 \quad R^2 = 0.057$$

- The model with lagged crime rate:

$$\widehat{\text{l\_crmrte87}} = \underset{(0.821)}{0.076} + \underset{(0.02)}{0.009}\,\text{unem87} - \underset{(0.109)}{0.140}\,\text{l\_lawexpc87} + \underset{(0.132)}{1.194}\,\text{l\_crmrte82}$$
$$n = 46 \quad R^2 = 0.680$$

- In the first model, crime rate decreases as unemployment increases. This is counterintuitive.
- After controlling for the crime rate in 1982 (5 years ago) coefficient on $unem$ is positive but insignificant.
- What is the elasticity of the current crime rate to the crime rate in the previous period?

Measurement Errors

- In some applications, it may be difficult or impossible to collect data on actual values of variables.
- If the true value is not observed (in other words we have an imprecise measure of a variable) then the observed value will contain measurement error.
- For example, income and consumption reported by households may be different than the actual values. They may tend to underreport their income level.
- In this section, we are interested in the properties of OLS estimators under measurement errors.
- We will examine measurement errors in two parts: (1) measurement errors in the dependent variable and (2) measurement errors in the explanatory variables.
- We will learn under what conditions measurement errors lead to inconsistency in OLS estimators.

Measurement Errors in the Dependent Variable

- Let y^* be the actual value of the dependent variable that we attempt to explain. For concreteness, suppose that y^* is the actual savings of households.

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- y is the observed (or reported) value. The difference between the observed value and the actual value is the measurement error in the population

$$e_0 = y - y^*$$

- From this we have $y^* = y - e_0$. Plugging this into the model we obtain:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u + e_0$$

- Now, the error term in the new model is $u + e_0$. Measurement error is now in the regression error term. Does OLS produce consistent estimators?

Measurement Errors in the Dependent Variable

- The model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \underbrace{u + e_0}$$

- If the measurement error, e_0, is uncorrelated with each x_j then consistent estimation is possible. If the measurement error is independent from explanatory variables then OLS estimators are unbiased and consistent.
- If the error term, u and the measurement error e_0 are independent (this is usually assumed), then we have:

$$\text{Var}(u + e_0) = \text{Var}(u) + \text{Var}(e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$$

- This means that measurement error in the dependent variable results in a larger error variance than when no error occurs.
- As a result, OLS estimators will have larger variances and standard errors. In this case, we may try to collect more "quality" data.

Measurement Errors in the Dependent Variable: Example

- Consider the following savings model:

$$sav^* = \beta_0 + \beta_1 inc + \beta_2 size + \beta_3 educ + \beta_4 age + u$$

  sav^*: actual household savings, sav: reported (observed) household savings, inc: annual household income, $size$: number of individuals in the household, $educ$: education level of the household head, age: age of the household head.
- When the measurement error $(sav - sav^*)$ creates a problem?
- We can assume that the measurement error is uncorrelated with income, size, education and age.
- On the other hand, we may think that families with higher incomes, or more education, report their savings more accurately.
- Since we cannot observe measurement error we may never be able to determine if the measurement error is correlated with income or education.

Measurement Error in Explanatory Variable

- Measurement error in x can lead to more serious problems than measurement errors in y.
- To determine conditions under which OLS estimators become inconsistent let us consider the simple regression model:

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

  Suppose that the first 4 Gauss-Markov assumptions hold.
- Here, x_1^* is the unobserved actual value and x_1 is the observed value.
- Then, the measurement error is

$$e_1 = x_1 - x_1^*$$

- Assume that the expected value of the measurement error is zero: $\mathsf{E}(e_1) = 0$

Measurement Error in Explanatory Variable

- Assume that the error term u is uncorrelated with both x_1^* and x_1 so that:

$$\mathsf{E}(y|x_1^*, x_1) = \mathsf{E}(y|x_1^*)$$

- This means that after controlling for x_1^* we no longer need x_1 in the model.
- If we use x_1 instead of x_1^*, what are the properties of OLS estimators? Are they still consistent?
- This depends on the assumption we make about the measurement error.
- There are two possible assumptions: (1) measurement error is uncorrelated with x_1.
- (2) measurement error is uncorrelated with unobserved actual value, x_1^*.

(1) e_1 and x_1 are uncorrelated

- This assumption can be written as

$$\mathsf{Cov}(x_1, e_1) = 0$$

- Since $e_1 = x_1 - x_1^*$, it must be the case that e_1 and x_1^* are correlated.
- Under this assumption, substituting $x_1^* = x_1 - e_1$ in the model we obtain:

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

- Expected value and variance of the composite error term:

$$\mathsf{E}(u - \beta_1 e_1) = 0, \quad \mathsf{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$$

- OLS estimators are consistent because the error term and x_1 are uncorrelated. But the variance will be higher.

(2) e_1 and x_1^* are uncorrelated (CEV Assumption)

- This is known as the "Classical Errors-in-Variables (CEV)". In the econometrics literature, when we talk about measurement error in explanatory variable we usually mean CEV.
- The CEV assumption can be written as:

$$\mathsf{Cov}(x_1^*, e_1) = 0$$

- The observed value can be written as the sum of actual value and measurement error:

$$x_1 = x_1^* + e_1$$

- Obviously, if x_1^* and e_1 are uncorrelated, then, x_1 and e_1 must be correlated:

$$\mathsf{Cov}(x_1, e_1) = \mathsf{E}(x_1 e_1) = \mathsf{E}(x_1^* e_1) + \mathsf{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$

- Under CEV assumption, the covariance between x_1 and e_1 is equal to the variance of the measurement error.

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- Recall that the model was written as:

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

- Since e_1 is included in the composite error term, its covariance with x_1 will create a problem.
- The covariance between composite error term and x_1 is

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

- Because this covariance is not 0, OLS estimators will be biased and inconsistent under CEV assumption
- We can calculate the amount of inconsistency in OLS.

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- In the simple regression model, the probability limit of the OLS estimator of the slope parameter is:

$$\begin{aligned}
\text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\
&= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\
&= \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right) \\
&= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)
\end{aligned}$$

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- Probability limit of the OLS estimator:

$$\text{plim}(\hat{\beta}_1) = \beta_1 \underbrace{\left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)}_{\leq 1} \neq \beta_1$$

- The term in the parenthesis will always be smaller than 1. If and only if $\sigma_{e_1}^2 = 0$ then it is 1.
- This means that: $\hat{\beta}_1$ is always closer to 0 than the true value β_1 is. This is called **attenuation bias**.
- If $\beta_1 > 0$ then $\hat{\beta}_1$ will approach a value smaller than the true value in the limit (underestimation). Otherwise, it will approach a bigger value (overestimation).

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- Probability limit of the OLS estimator:

$$\text{plim}(\hat{\beta}_1) = \beta_1 \underbrace{\left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)}_{\leq 1} \neq \beta_1$$

- If the variance of x_1^* is large as compared to the variance of e_1 then the ratio $\text{Var}(x_1^*)/\text{Var}(x_1)$ will be close to 1. In this case the amount of inconsistency may not be large. But it is almost impossible to determine this.
- Things are more complicated when we add more explanatory variables.
- But we can say that measurement errors generally lead to inconsistency of all OLS estimators.

(2) CEV Assumption: $\text{Cov}(x_1^*, e_1) = 0$

- Consider the following model for the college success:

$$colGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + u$$

$faminc$: Family income, $hsGPA$: high school GPA, SAT: Scholastic Aptitude Test result

- $faminc^*$ is the actual family income. If a questionnaire method is used to collect data then the student will be asked to report family income.
- We can collect data on hsGPA and SAT scores from student records. But we cannot do this for family income levels.
- If the reported income is different from the actual income, and if the CEV assumption is valid (ie actual income and measurement error are uncorrelated) then, OLS estimator for β_1 will be biased and inconsistent.
- As a result, the impact of the family income on the college success will be underestimated (downward bias).

Data Problems

- Measurement errors can be viewed as a data problem because we cannot obtain data on actual variables of interest.
- Another data problem that we saw before is multicollinearity among the explanatory variables. When two independent variables are highly correlated, it can be difficult to estimate the partial effect of each reflected by high standard errors. Remember that no assumption is violated in this case.
- There may be several other data problems:
- Missing data
- Nonrandom samples
- Outliers (extreme observations)

Missing Data

- The missing data problem can arise in a variety of forms. For example, in surveys respondent may not answer some of the questions.
- If data are missing for an observation on either the dependent variable or one of the independent variables, then the observation cannot be used in estimation. Econometric software packages usually ignore observations with missing data. As a result sample size decreases.
- Is there any serious statistical consequences of missing data? The answer depends on why the data are missing. If the data are missing at random then this does not cause any bias. The only result is that the sample is reduced and OLS estimates will be less precise.
- If the data are missing in a systematic way the OLS estimators may be biased. For example, in the birthweight example, if the probability that education is missing is higher for those people with lower than average level of education then we have systematic missing data. MLR.2 Random

Nonrandom Sampling

- Violation of MLR.2 Random Sampling. If the missing data results in nonrandom sample then we have a more serious problem.
- For example, in the wage equation, suppose we want to include IQ scores as an explanatory variable.
- If obtaining an IQ score is easier for those with higher IQs, then the sample is not representative of the population. Workers with high IQs will be over-represented in the sample.
- In this case MLR.2 may not hold and thus OLS estimators may be biased.

Nonrandom Sampling

- Certain types of nonrandom sampling do not cause bias or inconsistency.
- Sample can be chosen on the basis independent variables without causing any statistical problems.
- This is called **exogenous sample selection**.
- For example, consider the following saving equation:

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

- If our data set was based on a survey of people over 35 years of age, then we have exogenous sample selection, a type of nonrandom sampling.
- If the other assumptions are satisfied, then OLS is still unbiased and consistent. The reason is that conditional expectation

$$\mathrm{E}(saving|income, age, size)$$

is the same for any subset of the population described by ~~income, age, or size~~.

Nonrandom Sampling

- If the sample selection is based on the dependent variable, y, MLR. 2 will not be satisfied which will cause bias in OLS.
- This is called **endogenous sample selection**.
- Consider the following wealth equation:

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

- Suppose that only people with wealth below \$250,000 are included in the sample. This a kind of endogenous sample selection and will result in biased and inconsistent estimators.
- This is because the population regression

$$\mathrm{E}(wealth|educ, exper, age)$$

is not the same as the expected value conditional wealth being less than \$250,000.

Outliers - Influential Observations

- In some applications, (usually but not only in small data sets) the OLS estimators are sensitive to the inclusion of one or several observations
- An observation is an influential observation if dropping it from the analysis changes the key OLS estimates by a practically large amount.
- An outlier is an unusually large or small values in some observations.
- OLS can be sensitive to the outliers because in minimizing SSR, large residuals receive a lot of weight.
- How can we determine if an observation is outlier/influential observation?

Outliers - Influential Observations

- Outliers can occur for two reasons in practice: (1) a mistake has been made in collecting and entering the data (eg adding a zero by mistake or misplacing a decimal point), or (2) outlier is a feature of the distribution of the variable.
- In practical applications, it may be a good idea to examine summary statistics of variables, eg, mean, median, mode, minimum, maximum, standard deviation etc.
- It is not very clear what should be done if the outlier is a feature of the distribution.
- Outlying observations can provide important information by increasing the variation in the explanatory variables resulting in reduced standard errors.
- Usual practice is that OLS results are reported with and without outlying observations.

Outliers: Example

- Research and Development (R&D) intensity and firm performance:

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

$rdintens$: R&D expenditures as percentage of sales; $sales$: sales (in millions \$); $profmarg$: profits as a percentage of sales, %

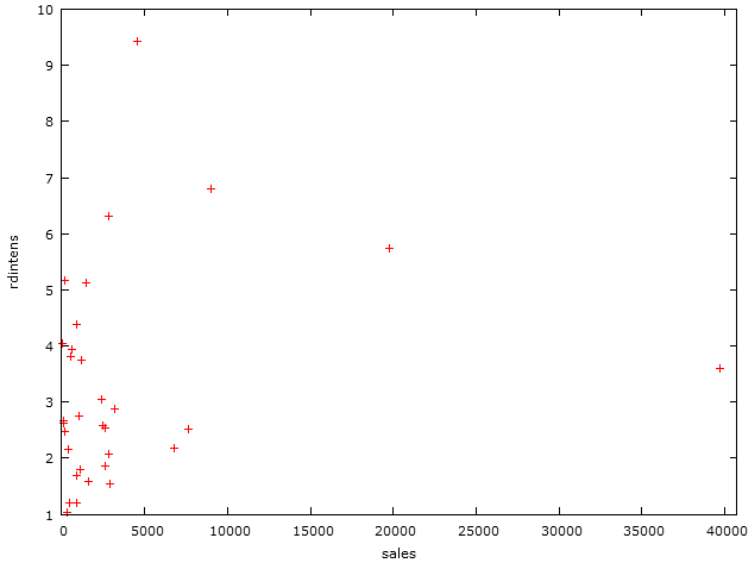
- Data set: RDCHEM.gdt, estimation results

$$\widehat{\text{rdintens}} = \underset{(0.585)}{2.62} + \underset{(0.00004)}{0.00005}\,\text{sales} + \underset{(0.046)}{0.045}\,\text{profmarg}$$

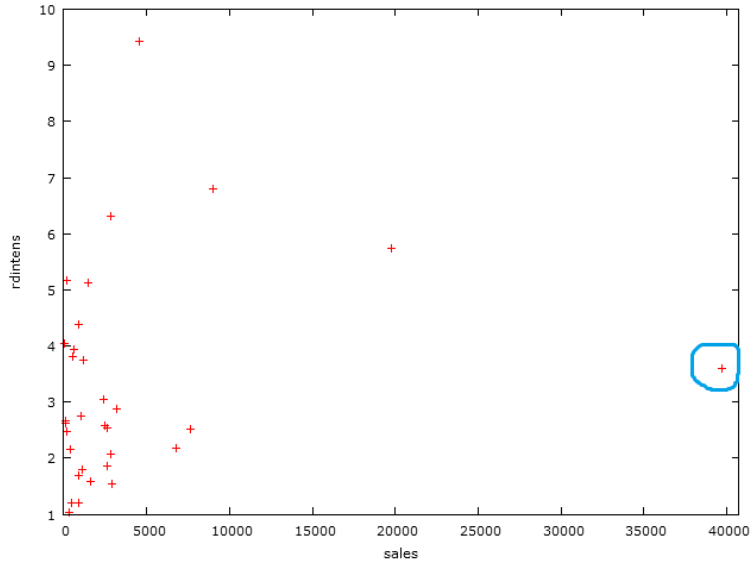
$$n = 32 \quad R^2 = 0.076$$

- Neither $sales$ nor $profmarg$ is statistically significant at even the 10% level.
- Are there any outliers? Let us examine the scatter diagram.

Outliers: Example



Outliers: Example



Outliers: Example

- Of the 32 firms, 31 have annual sales less than \$20 billion. One firm has annual sales of nearly \$40 billion
- This may be an outlier. Estimation results without outlier:

$$\widehat{\text{rdintens}} = \underset{(0.592)}{2.297} + \underset{(0.000084)}{0.000186}\,\text{sales} + \underset{(0.0445)}{0.0478}\,\text{profmarg}$$

$$n = 31 \quad R^2 = 0.1728$$

- When the largest firm is dropped from the regression, the coefficient on sales more than triples, and it now has a t statistic over 2.
- There is a statistically significant relationship between R&D intensity and sales.
- The profit margin is still insignificant and its coefficient has not changed much.

Outliers

- Certain functional forms may be less sensitive to outlying observations. Logarithmic transformation significantly narrows the range of the data that can potentially mitigate the problems created by outliers. For example, consider the following model

$$\log(rd) = \beta_0 + \beta_1 \log(sales) + \beta_2 profmarg + u$$

  rd: R$D expenditures, $millions
- $n = 32$ with outlier:

$$\widehat{\log(\text{rd})} = \underset{(0.468)}{-4.378} + \underset{(0.060)}{1.084}\,\log(\text{sales}) + \underset{(0.013)}{0.023}\,\text{profmarg}$$
$$n = 32 \quad R^2 = 0.918$$

- $n = 31$ without outlier:

$$\widehat{\log(\text{rd})} = \underset{(0.511)}{-4.404} + \underset{(0.067)}{1.088}\,\log(\text{sales}) + \underset{(0.013)}{0.0218}\,\text{profmarg}$$
$$n = 31 \quad R^2 = 0.9037$$

- Results are practically the same. Can we reject the null hypothesis of unit elasticity?