Specification and Data Issues: Part II

BS1802 Statistics and Econometrics

Jiahua Wu

382 Business School
j.wu@imperial.ac.uk

Imperial College Business School

Imperial means Intelligent Business

Roadmap

- Regression analysis with cross-sectional data
  - The multiple regression analysis
    - Basics: estimation, inference, analysis with dummy variables
    - More technically involved: asymptotics, heteroskedasticity, specification and data issues
- Advanced topics
  - Limited dependent variable models
  - Panel data analysis
  - Regression analysis with time series data

Outline (Wooldridge, Ch. 3.3, 3.4, 5.2, 9.2, 9.5)

- Model diagnostics
- Using proxy variables for unobserved x variables
- Outliers
- A possible model fitting strategy

Outline

- Model diagnostics
- Using proxy variables for unobserved x variables
- Outliers
- A possible model fitting strategy

Statistical Properties of OLS Estimators

Theorem (3.1)

With a "good" model, the OLS estimators are unbiased, i.e., $E(\hat{\beta}_j) = \beta_j$, $j = 0, 1, \ldots, k$

Theorem (4.1, Normal Sampling Distribution)

With a "good" model,

$$\hat{\beta}_j \sim Normal\left(\beta_j, Var(\hat{\beta}_j)\right),$$

where the variance is given by

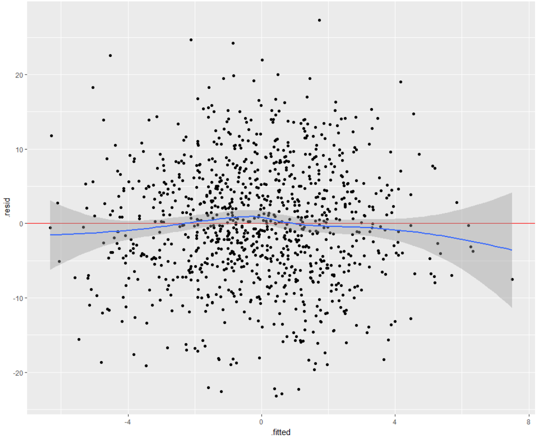
$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \qquad j = 1, \ldots, k.$$

Gauss-Markov Assumptions

- [MLR1] (linear in parameters) In the population model, y is related to x's by $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$, where $(\beta_0, \beta_1, \ldots, \beta_k)$ are population parameters and u is disturbance
  - Common causes lead to violation of this assumption
    - Functional form misspecification: log vs level form, omitting quadratic forms
  - Identification of assumption violation
    - Residual plots, RESET, comprehensive model selection, Davidson-MacKinnon test

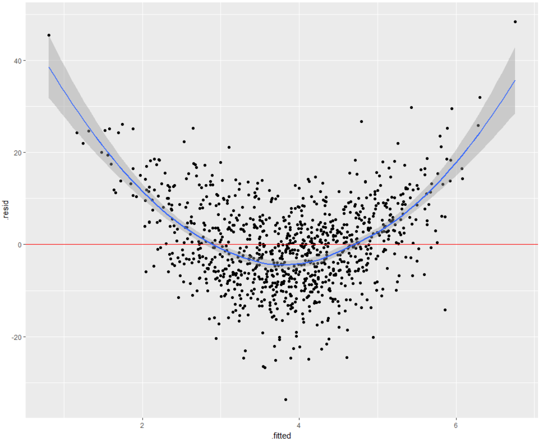
Residual Plots: A Correctly Specified Model

```
> x <- rnorm(1000, mean = 0, sd = 2)
> y <- x + rnorm(1000, mean = 0, sd = 8)
> m1 <- lm(y ~ x)
> ggplot(m1, aes(.fitted, .resid)) + geom_point() + geom_hline(
    yintercept=0, col="red") + stat_smooth(method = "loess")
```



Residual Plots: A Misspecified Model

```
> x <- rnorm(1000, mean = 0, sd = 2)
> y <- x + x^2 + rnorm(1000, mean = 0, sd = 8)
> m2 <- lm(y ~ x)
> ggplot(m2, aes(.fitted, .resid)) + geom_point() + geom_hline(
    yintercept=0, col="red") + stat_smooth(method = "loess")
```



Gauss-Markov Assumptions

- [MLR2] (random sampling) $\{(x_{i1}, \ldots, x_{ik}, y_i), i = 1, 2, \ldots, n\}$ with $n \geq k + 1$ is a random sample drawn from the population model

Missing Data

- If any observation is missing data on one of the variables in the model, it cannot be used.
- Would this practice cause problems?
  - If data is missing at random, then the only consequence is a reduction in the sample size
  - A problem can arise if the data is missing in a systematic way. The sample becomes nonrandom (violation to MLR2)
  - Eg. High income individuals are more likely to refuse to provide income data. This affects the "randomness" of sampling.

Nonrandom Samples

- Exogenous sample selection
  - If the sample is chosen on the basis of an explanatory variable x, the OLS estimators will still be unbiased
  - Eg. Consider the family savings model

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

    Suppose the data set is based on a survey of people aged 35 years and over
  - While the sample is nonrandom, zero-conditional mean assumption still holds as

$$E(u|\mathit{income}, \mathit{age}, \mathit{size}) = 0$$

    for any subset of $(\mathit{income}, \mathit{age}, \mathit{size})$

Nonrandom Samples

- Endogenous sample selection
  - If the sample is chosen on the basis of the dependent variable y, the OLS estimators will be biased
  - Eg. Consider the individual wealth model

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

    Suppose the data set is based on a survey of people with wealth below \$250,000
  - The sample is nonrandom and

    $$\begin{aligned} & E(wealth|educ, exper, age, wealth < 250,000) \\ \neq \quad & E(wealth|educ, exper, age) \end{aligned}$$

    Zero-conditional mean assumption fails!

Gauss-Markov Assumptions

- [MLR3] (no perfect collinearity) None of x's is constant and there is no perfect linear relationships among x's
  - Common causes lead to violation of this assumption
    - Multiple variables measure the same thing, dummy variables trap
  - Identification of assumption violation
    - Routinely reported by statistical softwares

A Model with Perfect Collinearity

```
> load("wage1.RData")
> male <- 1 - data$female
> wage.m1 <- lm(lwage ~ educ + exper + male + female, data)
```

```
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.137239   0.101327   1.354    0.176    
educ        0.091290   0.007123  12.816  < 2e-16 ***
exper       0.009414   0.001449   6.496 1.93e-10 ***
male        0.343597   0.037667   9.122  < 2e-16 ***
female            NA         NA      NA       NA    
```

Multicollinearity

- High correlation between two or more independent variables is known as multicollinearity
- Multicollinearity does not violate MLR3
- Effects of multicollinearity
  - Important variables can appear to be insignificant and standard errors can be large
  - Estimated coefficients can change substantially when variables are added or dropped

Multicollinearity: An Example (100 Observations)

```
> x1 <- rnorm(100, mean = 0, sd = 2)
> x2_1 <- x1 + rnorm(100, mean = 0, sd = 2)
> x2_2 <- x1 + rnorm(100, mean = 0, sd = 1)
> x2_3 <- x1 + rnorm(100, mean = 0, sd = 0.5)
> cor(x1, cbind(x1^2, x2_1, x2_2, x2_3))
```

```
                x2_1      x2_2      x2_3
[1,] -0.1341789 0.6799878 0.88835 0.9630341
```

```
> y <- x1 + rnorm(100, mean = 0, sd = 4)
> m1 <- lm(y ~ x1 + x2_1)
> m2 <- lm(y ~ x1 + x2_2)
> m3 <- lm(y ~ x1 + x2_3)
> stargazer(m1, m2, m3, align = TRUE, no.space = TRUE)
```

Multicollinearity: An Example (100 Observations)

	Dependent variable:		
	y		
	(1)	(2)	(3)
x1	0.968***	1.080***	2.722***
	(0.257)	(0.411)	(0.674)
x2_1	−0.072		
	(0.167)		
x2_2		−0.194	
		(0.377)	
x2_3			−1.887***
			(0.670)
Constant	−0.607*	−0.622*	−0.598*
	(0.351)	(0.353)	(0.338)
Observations	100	100	100
R^2	0.189	0.189	0.249
Adjusted R^2	0.172	0.173	0.233
Residual Std. Error (df = 97)	3.455	3.454	3.325
F Statistic (df = 2; 97)	11.285***	11.333***	16.054***
Note:			*p$<$0.1; **p$<$0.05; ***p$<$0.01

Multicollinearity: Another Example (1,000 Observations)

	Dependent variable:		
	y		
	(1)	(2)	(3)
x1	0.933***	1.067***	0.811***
	(0.090)	(0.143)	(0.276)
x2_1	0.087		
	(0.065)		
x2_2		−0.050	
		(0.128)	
x2_3			0.207
			(0.269)
Constant	−0.215*	−0.213*	−0.211
	(0.129)	(0.129)	(0.129)
Observations	1,000	1,000	1,000
R^2	0.205	0.203	0.204
Adjusted R^2	0.203	0.202	0.202
Residual Std. Error (df = 997)	4.077	4.080	4.079
F Statistic (df = 2; 997)	128.384***	127.356***	127.631***
Note:			*p<0.1; **p<0.05; ***p<0.01

Variance Inflation Factors

- The variance inflation factor for x_j is

$$VIF_j = \frac{1}{1 - R_j^2},$$

  R_j^2 is the R-squared from regressing x_j on all the other independent variables

  - x_j is strongly correlated with other independent variables $\rightarrow R_j^2$ close to $1 \rightarrow VIF_j$ is large
  - Rule of thumb: Value of VIF greater than 10 indicates the multicollinearity problem
  - R function: `vif` in multiple packages, such as `HH`, `car`, `fmsb`, `faraway` and `VIF`

Gauss-Markov Assumptions

- [MLR4] (zero conditional mean) The disturbance u satisfies $E(u|x_1, \ldots, x_k) = 0$ for any given value of $(x_1, \ldots, x_k)$
  - Common causes lead to violation of this assumption
    - Missing important variables in the model (either unobservable or fail to include them in the model)
  - Identification of assumption violation
    - Case-by-case: mostly based on intuition and subject knowledge
- MLR1-4 are required for OLS estimators to be unbiased.

Gauss-Markov Assumptions

- [MLR5] (homoskedasticity) $Var(u_i|x_{i1}, \dots, x_{ik}) = \sigma^2$ for $i = 1, 2, \dots, n$. (It implies $Var(u_i) = \sigma^2$)
  - Common causes lead to violation of this assumption
    - Data issue
  - Identification of assumption violation
    - Residual plots, Breusch-Pagan test, White test
  - Solutions
    - Robust standard errors
- MLR1-5 are collectively known as the Gauss-Markov Assumptions

Normality Assumption

- [MLR6] (normality) The disturbance u is independent of all explanatory variables and normally distributed with mean zero and variance σ^2:

$$u \sim \text{Normal}(0, \sigma^2)$$

- MLR1-6 imply the OLS estimators are normally distributed
- The normality leads to the exact distributions of the t stat and the F stat, which are the basis for inference

Large-Sample (Asymptotic) Inference

- MLR6 ($u \sim Normal(0, \sigma^2)$) is often too strong an assumption in practice
- How do we do inference without MLR6?
    - Central limit theorem (CLT) provides an answer
    - When n is large, the OLS estimators are approximately normally distributed

Large Sample (Asymptotic) Inference

- When u is not normally distributed, it is just as legitimate to write
$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1},$$
as t_{n-k-1} approaches $Normal(0, 1)$ for large $n - k - 1$
- For good approximation, how "large" must the n be?
  - Depends on the distribution of u.
- t testing, F testing and the construction of confidence intervals are carried out exactly the same as under Normality assumptions
- Note that while we no longer need to assume normality with a large sample, we do still need homoskedasticity

Model Diagnostics

- After fitting a regression model, it is important to determine whether all the necessary model assumptions are valid
- Any violations may invalidate subsequent inferential procedures, resulting in faulty conclusions

Outline

- Model diagnostics
- Using proxy variables for unobserved x variables
- Outliers
- A possible model fitting strategy

Unobserved Explanatory Variable

- What if model is misspecified because no data is available on an important x variable?
- Often the omitted-variable bias can be reduced by using a proxy variable
- A proxy variable must be related to the unobserved variable
- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$
  - (β_1, β_2) are parameters of interest and x_3^* is unobserved
  - But we have a proxy variable x_3, where $x_3^* = \delta_0 + \delta_3 x_3 + v_3$
  - Now suppose we just substitute x_3 for x_3^*
  - So, under what conditions will this solution give us unbiased estimates of β_1 and β_2?

Conditions for a Valid Proxy Variable

- A valid proxy (x_3) for a key unobserved variable (x_3^*):
  1. ZCM assumption holds for observed, unobserved and the proxy: $E(u|x_1, x_2, x_3, x_3^*) = 0$
  2. If x_3 is controlled for, the conditional mean of x_3^* does not depend on x_1 and x_2: $E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) = \delta_0 + \delta_3 x_3$
- That is, u is uncorrelated with x_1, x_2 and x_3^*, and v_3 is uncorrelated with x_1, x_2 and x_3
- Condition 2 implies $x_3^* = \delta_0 + \delta_3 x_3 + v_3$, and thus

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

- OLS are unbiased for estimating (β_1, β_2) under conditions 1 and 2

Conditions for a Valid Proxy Variable

- Without the conditions, we can end up with biased estimates
  - Say $x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3$
  - Substituting x_3^* into the model, we have

$$y = (\beta_0 + \beta_3 \delta_0) + (\beta_1 + \beta_3 \delta_1) x_1 + (\beta_2 + \beta_3 \delta_2) x_2 + \beta_3 \delta_3 x_3 + (u + \beta_3 v_3)$$

  - Bias will depend on signs of β_3 and δ_j
  - This bias may still be smaller than omitted variable bias, though
- In practice, lagged dependent variables are commonly used as proxies
  - to account for omitted variables that contribute to both past and current levels of y

Lagged Dependent Variables: An Example

- Example 9.4. City Crime Rates (crime2.RData)
  - Consider a simple equation to explain city crime rates

    $$crmrte = \beta_0 + \beta_1 unem + \beta_2 lawexpc + \beta_3 crmrte_{-1} + u,$$

    where

    - *crmrte*: a measure of per capita crime
    - *unem*: the city unemployment rate
    - *lawexpc*: per capita spending on law enforcement
    - $crmrte_{-1}$: the crime rate measured in some earlier year
  - The data are from 46 cities for the year 1987. The crime rate is also available for 1982.

Lagged Dependent Variables: An Example

Dependent Variable: log($crmrte_{87}$)

Independent Variables	(1)	(2)
$unem_{87}$	−.029 (.032)	.009 (.020)
log($lawexpc_{87}$)	.203 (.173)	−.140 (.109)
log($crmrte_{82}$)	—	1.194 (.132)
intercept	3.34 (1.25)	.076 (.821)
Observations *R*-squared	46 .057	46 .680

- Model (1): explanatory variables are insignificant with unexpected signs
- Model (2): use the lag of the dependent variable $crmrte_{82}$ as a proxy to control for unobserved factors

Outline

- Model diagnostics
- Using proxy variables for unobserved x variables
- Outliers
- A possible model fitting strategy

Outliers and Leverage

- Outliers are "unusual" observations that are far away from the "centre"
  - OLS is generally sensitive to outliers as large residuals once squared received much more weight in OLS
  - Outliers can be simple data entry errors
    - It is always a good idea to check summary statistics (min, max, etc)
    - Not unreasonable to fix observations where it's clear there was just an extra zero entered, etc.
  - Outliers can be that the observation is just truly very different from the others

Outliers and Leverage

- Outliers in the dependent variable
  - Observations lie far from the SRF
  - Rule of thumb: An outlier is an observation, whose residual is larger than 3 standard deviations away from the mean
- Outliers in the independent variable
  - Known as high-leverage points, to distinguish them from observations that are outliers in the response variable
  - Can be identified using leverage values or Cook's distances

Leverage Values

- Recall that
$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y.$$
That is, the fitted values of a multiple regression can be written as
$$\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \cdots + p_{in}y_n$$
- p_{ii} is called the leverage value for the ith observation
  - It measures the "outlierness" in the independent variables
  - $0 \leq p_{ii} \leq 1$, and average of all leverage values is $(k+1)/n$
  - Rule of thumb: Points with p_{ii} greater than $2(k+1)/n$ are generally regarded as points with high leverage
  - Points with high leverage should be flagged and examined

Cook's Distance

- Cook's distance measures the influence of the jth observation by

$$C_j = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_{i(j)})^2}{\hat{\sigma}^2(k+1)},$$

where $\hat{y}_i$ is the fitted value obtained from the full sample, and $\hat{y}_{i(j)}$ is the fitted value obtained by deleting the jth observation

  - Cook's distance can be thought of as the product of leverage and outlierness
  - If a point is influential, its deletion causes large changes in fitted values, and value of C_j will be large
  - Rule of thumb: Points with C_j values greater than 1 are influential

Outliers: An Example

- Example 9.8. R&D Intensity and Firm Size (rdchem.RData)
  - The regression model is

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

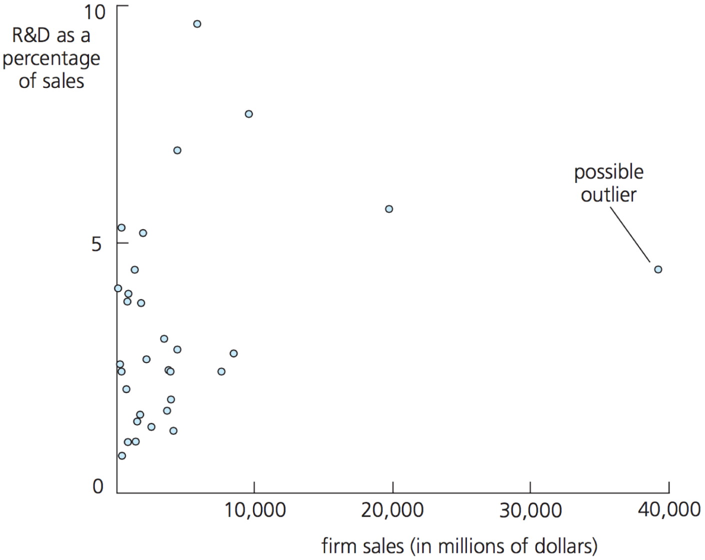
    where

    - *rdintens*: R&D expenditures as a percentage of sales
    - *sales*: annual sales (in millions)
    - *profmarg*: profits as a percentage of sales
  - The OLS equation using data on 32 chemical companies is

$$\widehat{rdintens} = \underset{(0.586)}{2.625} + \underset{(.000044)}{.000053} sales + \underset{(.0462)}{.0446} profmarg,$$

    $n = 32, R^2 = .0761$
  - Neither *sales* nor *profmarg* is statistically significant at even the 10% level in this regression.

Outliers: An Example



Outliers: An Example

- Of the 32 firms, 31 have annual sales less than \$20 billion, where one firm has annual sales of almost \$40 billion.
- Without the high-leverage observation, the estimated model is given by

$$\widehat{rdintens} = \underset{(0.592)}{2.297} + \underset{(.000084)}{.000186} sales + \underset{(.0445)}{.0478} profmarg,$$

$n = 31, R^2 = .173$

  - Using the sample of smaller firms, there is a statistically significant positive effect between R&D intensity and firm size.
  - The profit margin is still not significant, and its coefficient has not changed by much.

Outline

- Model diagnostics
- Using proxy variables for unobserved x variables
- Outliers
- A possible model fitting strategy

A Possible Model Fitting Strategy

1. Understand the data set
   - Examine the variables $y, x_1, x_2, \ldots, x_k$ one at a time; Calculate the summary statistics, and also graphically by looking at histograms or box plots
   - Construct pairwise scatter plots
2. Regression and variable selection
   - Model selection based on $\bar{R}^2$ and information criteria
   - Test for correct functional forms of variables

A Possible Model Fitting Strategy

3. Residual analysis - ensure satisfactory residual plots and no negative diagnostic messages. If needed, repeat Step 2.
    - Check linearity. If none, make a transformation on the variable
    - Check for heteroscedasticity
    - Look for outliers and high-leverage points
4. Model validation
    - The model may be fitted by part of the data and validated by the remainder of the data when the amount of data is large
    - Otherwise, resampling methods such as bootstrap, jackknife and cross-validation can be used