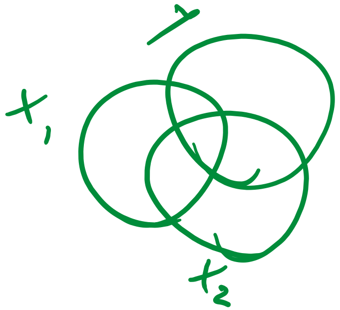
Class 14- Multiple Regression Model Estimation (Part IV)

Pedram Jahangiry

Variances in misspecified models



The choice of whether to include a particular variable in a regression can be made by analyzing the tradeoff between bias and variance

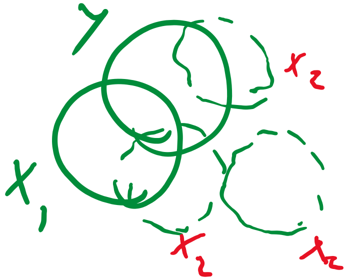
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$ ← True population model (UNOBSERVED)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$ ← Estimated model 1

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$ ← Estimated model 2

It might be the case that the likely omitted variable bias in the misspecified model 2 is overcompensated by a smaller variance

Variances in misspecified models (cont.)



2

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

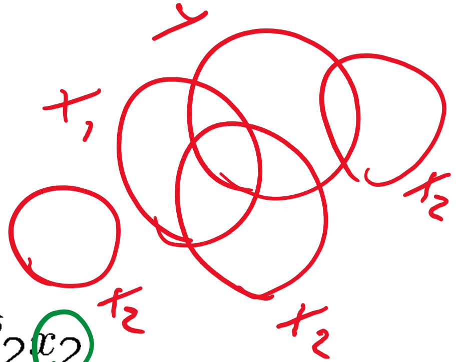
$$Var(\tilde{\beta}_1) = \sigma^2 / SST_1$$

1

^

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$Var(\hat{\beta}_1) = \sigma^2 / \left[SST_1 (1 - R_1^2) \right]$$



* Three cases:

- When x_1 and x_2 are uncorrelated: ✓ $\widetilde{\beta_1}$ *and* $\widehat{\beta_1}$ are both unbiased, ✓ and $\text{var}(\widetilde{\beta_1}) \geq var(\widehat{\beta_1})$
- When x_1 and x_2 are correlated and:
  - When $\beta_2 = 0$: ✓ $\widetilde{\beta_1}$ *and* $\widehat{\beta_1}$ are both unbiased, ✓ and $\text{var}(\widetilde{\beta_1}) < var(\widehat{\beta_1})$
  - When $\beta_2 \neq 0$: ✓ $\widetilde{\beta_1}$ is biased, $\widehat{\beta_1}$ is unbiased, ✓ and $\text{var}(\widetilde{\beta_1}) < var(\widehat{\beta_1})$

- ✓ ✓ Trade off between bias and variance
- ✓ ✓ Variance will shrink as sample size gets larger (multicollinearity becomes less important as n gets larger)
- ✓ ✓ Bias will not vanish even in large samples

$\hat{\sigma}^2$

Estimating the error variance

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2\right) / \underbrace{[n-k-1]}_{df}$$

estimated params.

$df = n-k-1 = n-\underbrace{(k+1)}$

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients $(k+1)$ from the number of observations (n). The number of observations minus the number of estimated parameters is also called the degrees of freedom.

$$\boxed{E(\hat{\sigma}^2) = \sigma^2}$$

THEOREM 3.3	UNBIASED ESTIMATION OF σ^2
	Under the Gauss-Markov assumptions MLR.1 through MLR.5, $E(\hat{\sigma}^2) = \sigma^2$.

Estimation of the sampling variances of the OLS estimators

The **true sampling** variation of the estimated β_j

$$sd(\widehat{\beta}_j) = \sqrt{Var(\widehat{\beta}_j)} = \sqrt{\sigma^2 / \left[SST_j(1 - R_j^2)\right]}$$

Plug in $\widehat{\sigma}^2$ for the unknown σ^2

The **estimated sampling** variation of the estimated β_j

$$se(\widehat{\beta}_j) = \sqrt{\widehat{Var}(\widehat{\beta}_j)} = \sqrt{\widehat{\sigma}^2 / \left[SST_j(1 - R_j^2)\right]}$$

Note that these formulas are only valid under assumptions MLR.1-MLR.5 (there has to be homoskedasticity)

MRM in R

```r
library(wooldridge)

MRM <- lm(wage ~ educ + exper , wage1)
summary(MRM)
```

```
Call:
lm(formula = wage ~ educ + exper, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5532 -1.9801 -0.7071  1.2030 15.8370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.39054    0.76657  -4.423 1.18e-05 ***
educ         0.64427    0.05381  11.974  < 2e-16 ***
exper        0.07010    0.01098   6.385 3.78e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.257 on 523 degrees of freedom
Multiple R-squared:  0.2252,	Adjusted R-squared:  0.2222 
F-statistic: 75.99 on 2 and 523 DF,  p-value: < 2.2e-16
```

```r
library(wooldridge)
library(stargazer)

MRM <- lm(wage ~ educ + exper , wage1)
stargazer(MRM, type = "text")
```

```
===============================================
                        Dependent variable:    
                    ---------------------------
                               wage            
-----------------------------------------------
educ                         0.644***          
                              (0.054)          

exper                        0.070***          
                              (0.011)          

Constant                     -3.391***         
                              (0.767)          

-----------------------------------------------
Observations                    526            
R2                             0.225           
Adjusted R2                    0.222           
Residual Std. Error      3.257 (df = 523)      
F Statistic           75.990*** (df = 2; 523)  
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

526 - 2 - 1 = 523

Efficiency of OLS: The Gauss-Markov Theorem

- Under assumptions MLR.1 - MLR.4, OLS is unbiased
- However, under these assumptions there may be many other estimators that are unbiased
- Which one is the unbiased estimator with the smallest variance?
- In order to answer this question one usually limits oneself to **linear estimators**, i.e. estimators linear in the dependent variable

$Y, X_1, X_2, \ldots \rightarrow$ OLS $\rightarrow \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots$

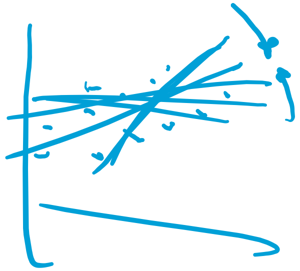
MLE, GLS, WLS, FGLS, ⋮ $\rightarrow \tilde{\beta}$

MLR.1 - MLR.5 → OLS is BLUE

Best, linear

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$$

May be an arbitrary function of explanatory variables; the OLS estimator can be shown to be of this form



THEOREM 3.4

GAUSS-MARKOV THEOREM

Under Assumptions MLR.1 through MLR.5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the best linear unbiased estimators (BLUEs) of $\beta_0, \beta_1, \ldots, \beta_k$, respectively.

Best

OLS

others

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j) \quad j = 0, 1, \ldots, k$$

$$\text{for all} \quad \tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i \quad \text{for which} \quad E(\tilde{\beta}_j) = \beta_j, j = 0, \ldots, k$$

- ❑ OLS is only the best estimator if MLR.1 – MLR.5 hold
- ❑ If there is heteroskedasticity for example, there are better estimators with lower variances

~~MLR5~~

WLS

THE GAUSS-MARKOV ASSUMPTIONS

The following is a summary of the five Gauss-Markov assumptions that we used in this chapter. Remember, the first four were used to establish unbiasedness of OLS, whereas the fifth was added to derive the usual variance formulas and to conclude that OLS is best linear unbiased.

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

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

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the independent variables. In other words,

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0.$$

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variables. In other words,

$$\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2.$$

Comments on the Language of MRM

- OLS is an estimation method, not a model. So it is WRONG to say that "I estimated an OLS model"
- Proper way to introduce a discussion of the estimates is to say:

I estimated the following equation by ordinary least squares.

$$\begin{aligned} math4 = \beta_0 + \beta_1 classize4 + \beta_2 math3 + \beta_3 \log(income) \\ + \beta_4 motheduc + \beta_5 fatheduc + u \end{aligned} \quad + \beta_6 X_6$$

- ✓ First I argue whether it is reasonable to maintain Assumption MLR.4, by focusing on the factors that might still be in u
- ✓ Then under the assumption that no important variables have been omitted from the equation, and assuming random sampling, the OLS estimator is unbiased.
- ✓ If the error term u has constant variance, the OLS estimator is actually best linear unbiased."

MLR5 $\rightarrow$