

Class notes week 4

OUTLINE

(Un)biasedness of OLS

- Violations of zero conditional mean assumption
- Violations of perfect collinearity

Efficiency of OLS and Gauss-Markov assumptions

Problems causing large(r) standard errors

- Multicollinearity
- Misspecification

What should you learn from this class?

- Recognize violations of the Gauss-Markov assumptions
- Know consequences of violations
- Understand multicollinearity problem
- Apply this understanding when interpreting regression results

Back to the Broad Picture: the “WHAT”

We are interested in understanding the effect of a variable x on variable y

- ⇒ Need a coefficient estimate
 - Direction
 - Magnitude
- ⇒ Need to know how precise this estimate is

Luckily, if the data is “well behaved”, minimizing the sum of squared residuals will give us a “good” estimate of the coefficient and its precision

More precise:

- If the data satisfy 4 assumptions than OLS gives us an unbiased estimate of the coefficient
- If the data satisfy 4+1 assumptions, than OLS gives us an unbiased estimate of the variance of the coefficient estimate and OLS is efficient
- If the data satisfy 4+2 assumptions, than the coefficients have a normal distribution

Back to the Broad Picture: the “WHY”

- Read and interpret OLS regression results
 - o What does the coefficient mean?
 - o What does the standard error stand for?
 - o What does R^2 stand for?
 - o ...
- Know which assumptions are underlying the OLS estimates
 - o Critically evaluate whether these assumptions are reasonable for a particular problem and data set
 - o Understand what deviations from assumptions mean for the interpretation of the OLS results
 - o Think about how to avoid violations of the assumptions

4 Assumptions for Unbiasedness of OLS

1) Linearity

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

2) Random Sampling

$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i=1, \dots, n\}$: random sample from the population

3) Zero Conditional Mean

$$E(u|x_1, x_2, \dots, x_k)=0$$

4) No Perfect Collinearity

No exact *linear* relationships among the independent variables

Violations of Zero Conditional Mean Assumption

- omitted variable
- measurement error in x
- x and y jointly determined
(extreme: y determines x – reverse causality)

Terminology:

- Independent variable is not *exogenous*
- *Endogeneity* problem
- Problem of *identification*

Violations of no perfect collinearity

⇒ Can't write one explanatory variable as a linear combination of the other explanatory variables

e.g. Assumption is violated if there exists (a, b) such that $x_1 = a + bx_2$

- ⇒ One variable can't be multiple from another
- ⇒ One variable can't be the sum of some of the others
- ⇒ When variables are shares: can't include all the shares

Practical Note:

- Stata will not estimate models with perfect collinearity
- But watch out for interpretations!

Example 1: Examples of perfect collinearity - Voting outcomes and campaign expenditures

Source: Wooldridge, VOTE1.dta (From M. Barone and G. Ujifusa, *The Almanac of American Politics*, 1992. Washington, DC: National Journal.) – two-party races for the US House of Representatives in 1988.

Variable description:

voteA	byte	%5.2f	percent vote for A
expendA	float	%8.2f	camp. expends. by A, \$1000s
expendB	float	%8.2f	camp. expends. by B, \$1000s
shareA	float	%5.2f	$100 * (\text{expendA} / (\text{expendA} + \text{expendB}))$
shareB	float	%5.2f	$100 * (\text{expendB} / (\text{expendA} + \text{expendB}))$

Variable	Obs	Mean	Std. Dev.	Min	Max
voteA	173	50.50289	16.78476	16	84
expendA	173	310.611	280.9854	.302	1470.674
expendB	173	305.0885	306.2783	.93	1548.193
shareA	173	51.07654	33.48358	.094635	99.495
shareB	173	48.92346	33.48358	.5049973	99.90536

Relationship between % of vote for party A and % expenditures by party A & B

. regress voteA shareA shareB

Source	SS	df	MS	Number of obs =	173
Model	41486.2306	1	41486.2306	F(1, 171) =	1017.66
Residual	6971.01793	171	40.7661867	Prob > F =	0.0000
				R-squared =	0.8561
				Adj R-squared =	0.8553
Total	48457.2486	172	281.728189	Root MSE =	6.3848

voteA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
shareA	(dropped)				
shareB	-.4638269	.0145397	-31.90	0.000	-.4925272 - .4351266
_cons	73.19491	.8611812	84.99	0.000	71.49499 74.89482

```
. regress voteA shareB
```

Source	SS	df	MS			
Model	41486.2306	1	41486.2306	Number of obs =	173	
Residual	6971.01793	171	40.7661867	F(1, 171) =	1017.66	
Total	48457.2486	172	281.728189	Prob > F =	0.0000	
				R-squared =	0.8561	
				Adj R-squared =	0.8553	
				Root MSE =	6.3848	

voteA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
shareB	-.4638269	.0145397	-31.90	0.000	-.4925272	-.4351266
_cons	73.19491	.8611812	84.99	0.000	71.49499	74.89482

```
. regress voteA shareA
```

Source	SS	df	MS			
Model	41486.2307	1	41486.2307	Number of obs =	173	
Residual	6971.01783	171	40.7661862	F(1, 171) =	1017.66	
Total	48457.2486	172	281.728189	Prob > F =	0.0000	
				R-squared =	0.8561	
				Adj R-squared =	0.8553	
				Root MSE =	6.3848	

voteA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
shareA	.4638269	.0145397	31.90	0.000	.4351266	.4925272
_cons	26.81221	.8872146	30.22	0.000	25.06091	28.56352

Interpretation: the variables shareA and shareB are perfectly collinear (shareA = 100-shareB). Therefore they cannot both be used as independent variables in the regression. If you try to add them both, STATA will automatically drop one. Comparing the first two estimates you see that gives the same result than estimating it with only shareB. Comparing the estimates with shareA as the only independent variable, and with shareB as the only independent variable we see they give us the same information: Increasing the share of expenditures of B with one percentage point (which implies a one percentage point decrease of the share of A, is predicted to decrease the share of votes for A with .46 percentage points, ceteris paribus.

5 Gauss-Markov Assumptions (4+1)

- 1) Linearity
- 2) Random Sampling
- 3) Zero Conditional Mean
- 4) No Perfect Collinearity
- 5) Homoscedasticity

$$\text{Var}(u | x_1, \dots, x_k) = \sigma^2$$

Variance in the error term, conditional on the explanatory variables, is constant

- ⇒ OLS estimate of error variance is unbiased
- ⇒ Formula for sampling variance of the OLS coefficients
- ⇒ OLS is efficient
(i.e. variance is the smallest variance possible)

Unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{(n-k-1)} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{(n-k-1)}$$

Why n-k-1?

k+1: number of conditions imposed by
minimizing sum of squares (OLS)

n-k-1 = degrees of freedom

= number of observations - number of estimated parameters

Assumption 1 to 5

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

Sampling variance of the OLS coefficients

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

⇒ More precise estimate when:

- Lower variance of the error term
("less noise")
- Higher variation in the x_j 's : $\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$
- Less correlation between x_j 's
(R_j^2 is a measure of correlation between independent variables)

⇒ Problem of *multicollinearity*

Note: standard deviation of $\hat{\beta}_j$:

$$sd(\hat{\beta}_j) = \text{Var}(\hat{\beta}_j)^{1/2}$$

⇒ Standard error of $\hat{\beta}_j$:

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{[(1-R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2]^{1/2}}$$

R^2 versus $\text{Var}(\hat{\beta}_j)$

$$R^2 \equiv \frac{SSE}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$\text{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \hat{u}_i^2}{(n - k - 1) * \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 * (1 - R_j^2)}$$

Intuition:

R^2 = goodness of fit
= % of sample variation that is explained by all the explanatory variables together

$\text{Var}(\hat{\beta}_j)$ = Measure of precision of the coefficient of one explanatory variable

Efficiency of OLS: Gauss-Markov Theorem

If assumptions 1 to 5 are satisfied, the OLS gives us the Best Linear Unbiased Estimators (BLUE)

- Best?
⇒ OLS has the smallest variance, i.e.
the OLS estimators are more precise than other potential estimators
- Linear?
⇒ $\hat{\beta}_j$ can be written as a linear combination of y_i 's
- Unbiased?
⇒ We know because of assumptions 1 to 4

Multicollinearity

Not a violation of any assumption!
⇒ OLS is still BLUE

BUT: Causes large standard errors
⇒ Statistical inference is more difficult

What can be done about this?

- Collect more data
- Redefine research question
- Drop a variable?
 - ⇒ BUT: might lead to omitted variable bias

Example of Multicollinearity : Relationship between education and family background

(see also tutorial 2)

Source: WAGE2.dta, Wooldridge, (data used in M. Blackburn and D. Neumark (1992), “Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials,” *Quarterly Journal of Economics* 107, 1421-1436.)

. desc

variable name	storage type	display format	value label	variable label
educ	byte	%9.0g		years of education
sibs	byte	%9.0g		number of siblings
meduc	byte	%9.0g		mother's education
feduc	byte	%9.0g		father's education
brthord	byte	%9.0g		birth order

. sum

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	935	13.46845	2.196654	9	18
sibs	935	2.941176	2.306254	0	14
meduc	857	10.68261	2.849756	0	18
feduc	741	10.21727	3.3007	0	18
brthord	852	2.276995	1.595613	1	10

. regress educ sibs meduc

Source	SS	df	MS	Number of obs =	857
Model	627.549185	2	313.774593	F(2, 854) =	76.43
Residual	3505.84638	854	4.10520653	Prob > F =	0.0000
				R-squared =	0.1518
				Adj R-squared =	0.1498
Total	4133.39557	856	4.82873314	Root MSE =	2.0261

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sibs	-.1403635	.0319623	-4.39	0.000	-.2030974 - .0776296
meduc	.248872	.0253694	9.81	0.000	.1990784 .2986656
_cons	11.32153	.317941	35.61	0.000	10.6975 11.94557

```
. regress educ sibs meduc feduc
```

Source	SS	df	MS	Number of obs = 722		
Model	772.281437	3	257.427146	F(3, 718)	=	65.20
Residual	2834.93324	718	3.94837499	Prob > F	=	0.0000
				R-squared	=	0.2141
				Adj R-squared	=	0.2108
Total	3607.21468	721	5.00307168	Root MSE	=	1.9871

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sibs	-.0936359	.0344713	-2.72	0.007	-.1613124	-.0259594
meduc	.1307872	.032689	4.00	0.000	.0666098	.1949646
feduc	.2100041	.0274748	7.64	0.000	.1560635	.2639447
_cons	10.36426	.3585001	28.91	0.000	9.660422	11.06809

Interpretation: the number of siblings, and mother's education is likely correlated with father's education (this is confirmed by the correlation coefficients below). Since father's education is also correlated with education of his child, omitting father's education from the first regression resulted in omitted variable bias. (positive bias of the effect of mother's education, negative bias for sibs). Given that the correlation between meduc and feduc is high, including father's education results however in multicollinearity. As a result, the standard error of the coefficient of meduc increased substantially. Given that multicollinearity is not a violation of any assumption, we prefer the second over the first estimation.

```
. corr sibs meduc feduc
(obs=722)
```

	sibs	meduc	feduc
sibs	1.0000		
meduc	-0.2913	1.0000	
feduc	-0.2269	0.5765	1.0000

```
. gen avpareduc = (meduc + feduc)/2
. regress educ sibs avpareduc
```

Source	SS	df	MS	Number of obs = 722		
Model	763.455959	2	381.72798	F(2, 719)	=	96.51
Residual	2843.75872	719	3.95515817	Prob > F	=	0.0000
				R-squared	=	0.2116
				Adj R-squared	=	0.2095
Total	3607.21468	721	5.00307168	Root MSE	=	1.9888

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sibs	-.0892415	.0343752	-2.60	0.010	-.1567293	-.0217538
avpareduc	.3496428	.0283853	12.32	0.000	.2939149	.4053708
_cons	10.23665	.3484904	29.37	0.000	9.552468	10.92083

Interpretation: I have now redefined the research question, and instead look at the effect of the average education level of the parents. Interestingly, we see that the effect is larger than for father's or mother's education separately.

Note: No need to do anything if control variable
i.e. if x_1 is the variable of interest, I don't
care about whether x_2 and x_3 are correlated

Misspecification

Including irrelevant variables in model

=> no bias

⇒ BUT: might increase the variance if extra variable is correlated with another independent variable

Source	SS	df	MS	Number of obs = 663		
Model	692.455912	4	173.113978	F(4, 658) =	43.75	
Residual	2603.75525	658	3.95707485	Prob > F =	0.0000	
				R-squared =	0.2101	
				Adj R-squared =	0.2053	
Total	3296.21116	662	4.97917094	Root MSE =	1.9892	

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sibs	-.0910052	.0426604	-2.13	0.033	-.1747722	-.0072383
meduc	.1214772	.0343839	3.53	0.000	.0539617	.1889926
feduc	.2152426	.0288683	7.46	0.000	.1585576	.2719277
brthord	-.0122032	.0647301	-0.19	0.851	-.1393056	.1148992
_cons	10.43929	.392977	26.56	0.000	9.667645	11.21093

regress educ sibs meduc feduc if brthord!=.

Source	SS	df	MS	Number of obs = 663		
Model	692.315273	3	230.771758	F(3, 659) =	58.40	
Residual	2603.89589	659	3.95128359	Prob > F =	0.0000	
				R-squared =	0.2100	
				Adj R-squared =	0.2064	
Total	3296.21116	662	4.97917094	Root MSE =	1.9878	

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sibs	-.0953664	.0358173	-2.66	0.008	-.1656961	-.0250367
meduc	.1221488	.0341738	3.57	0.000	.0550461	.1892515
feduc	.2155921	.0287876	7.49	0.000	.1590655	.2721186
_cons	10.41426	.3696027	28.18	0.000	9.688516	11.14

Interpretation: The first regression suggests that birthorder has no significant effect on education. There has to be however a correlation between the number of siblings and the birthorder, which causes some multicollinearity. As a result the standard error for the coefficient of sibs in the first model is larger than in the second model.