POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

Outline:

1. Underspecifying vs. Overspecifying a Model
2. Null hypothesis *t* tests
3. Variances and Standard Errors of Regression Coefficients
4. Variance Inflation Factors and Multicollinearity
5. Example: *Judicial Independence and Human Rights*

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

1. Underspecifying vs. Overspecifying a Model [3.3]

Underspecifying is omitting a relevant variable (i.e., in multiple regression $\widehat{\beta_j} \neq 0$)

The bad news about underspecifying: all coefficients are likely to be biased.

The good news about underspecifying: there is no good news.

Overspecifying is including an irrelevant variable (i.e., in multiple regression $\widehat{\beta_j} = 0$).

The bad news about overspecifying: standard errors are biased upward.

The good news about overspecifying: estimated coefficients are unbiased.

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

2. Null hypothesis *t* tests

Null hypothesis: $H_{0j} : \beta_j = 0$

Alternative hypothesis: $H_{1j} : \beta_j \neq 0$

Test statistic: $t = \dfrac{\widehat{\beta_j}}{se(\widehat{\beta_j})}$

Null hypothesis: $H_{0j} : \beta_j = a_j$

Alternative hypothesis: $H_{1j} : \beta_j \neq a_j$

Test statistic: $t = \dfrac{(estimate - hypothesized\ value)}{standard\ error\ of\ estimate} = \dfrac{(\widehat{\beta_j} - a_j)}{se(\widehat{\beta_j})}$

The *t* statistic answers the question: How many standard errors separate the estimated coefficient ($\widehat{\beta_j}$) from the null hypothesis (0 or a_j) for the coefficient?

Values of the *t* statistic sufficiently far from 0 result in a rejection of H_{0j} ;
our **rejection rule** is to reject H_{0j} if and only if $\left| t_{\widehat{\beta_j}} \right| > c$;
by convention, *c* is chosen to make areas above $+c$ and below $-c$ sum to .05

If H_{0j} is rejected at the 5% level, then we conventionally say "x_j is statistically significant."

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

3. Variances and Standard Errors of Regression Coefficients [3.4 , 5.2]

In addition to assuming linearity [MLR.1], simple random sampling [MLR.2], and no perfect multicollinearity among regressors[MLR.3], we also require :

[MLR.4] Zero conditional mean disturbance $E(u \mid x_1, x_2, \ldots, x_k) = 0$

[MLR.5] Homoskedastic disturbances $Var(u \mid x_1, x_2, \ldots, x_k) = \sigma^2$

Collectively, these five assumptions are known as the ***Gauss-Markov Assumptions***.

A stronger version of the last two assumptions is:

[MLR.6] Normality of disturbances $u \sim N(0, \sigma^2)$

which implies that the disturbances are independent of regressors and homoskedastic.

Fortunately, this stronger assumption is unnecessary, thanks to the Central Limit Theorem (and the random sampling assumption). Instead, we assume <u>asymptotic</u> normality, which tells us that the $\widehat{\beta}_j$ are approximately normally distributed around the true β_j. [Theorem 5.2]

Appendix 5A has a sketch of the proof of asymptotic normality.

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

3. Variances and Standard Errors of Regression Coefficients [3.4 , 5.2]

Recall from ch. 2:

$\hat{\sigma}^2 = \frac{1}{n-2}\sum \hat{u}_i^{\,2}$ was the estimated error variance [2.61]

$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum \hat{u}_i^{\,2}}{n-2}}$ was the standard error of a regression [2.62]

For a multiple regression model, we replace these with:

$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum \hat{u}_i^{\,2}$ is the estimated error variance [3.56]

$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum \hat{u}_i^{\,2}}{n-k-1}}$ is the standard error of the regression

p. 100 explains degrees of freedom in terms of the restrictions imposed on the residuals ($\hat{u}_i$);
$df = n - (k + 1)$ because there are $k + 1$ equations: $\sum \hat{u}_i = 0$

$\sum \hat{u}_i x_{ji} = 0$ for each $j = 1, \ldots, k$

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

3. Variances and Standard Errors of Regression Coefficients [3.4 , 5.2]

Recall from ch. 2:

$$var(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}$$ was the sampling variance of slope (β) [2.57]

$$se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$$ was the standard error of the slope (β) (p. 56)

For a multiple regression model, we replace these with:

$$var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sum(x_{ji} - \overline{x_j})^2 \cdot \left[1 - R_j^2\right]}$$ is sampling variance of a coefficient $\hat{\beta}_j$ [3.54]

where R_j^2 is coefficient of determination from regressing x_j on all $x_{\sim j}$

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\sum(x_{ji} - \overline{x_j})^2 \cdot \left[1 - R_j^2\right]}}$$ is the standard error of a coefficient $\hat{\beta}_j$ [3.58]

$$= \frac{\hat{\sigma}}{\sqrt{n} \cdot s_{x_j} \cdot \sqrt{1 - R_j^2}}$$ where s_{x_j} is the s.d. of x_j [3.59]

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

3. Variances and Standard Errors of Regression Coefficients [3.4 , 5.2]

Let us look closer at the standard error of coefficient $\widehat{\beta_j}$:

$$se(\widehat{\beta_j}) = \frac{\hat{\sigma}}{\sqrt{\sum(x_{ji}-\overline{x_j})^2 \cdot \left[1-R_j^2\right]}} = \frac{\hat{\sigma}}{\sqrt{SST_j}\cdot\sqrt{1-R_j^2}} = \frac{\hat{\sigma}}{\sqrt{n}\cdot s_{x_j}\cdot\sqrt{1-R_j^2}}$$

The three elements of the standard error of a coefficient $\widehat{\beta_j}$ are:

(a) $\hat{\sigma}$ the standard error of the regression

(b) $\sqrt{\sum(x_{ji}-\overline{x_j})^2} = \sqrt{n}\cdot s_{x_j}$ where s_{x_j} is the standard deviation of x_j

(c) $\sqrt{1-R_j^2}$ where R_j^2 is the R^2 from regression of x_j on all other x's

Adding a control or proxy variable decreases $\hat{\sigma}$ but also potentially decreases $\sqrt{1-R_j^2}$.

Enlarging the sample (n) increases $\sqrt{n}\cdot s_{x_j}$, which decreases $se(\widehat{\beta_j})$.

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

4. Variance Inflation Factors and Multicollinearity

Some 'casual' methods for detecting multicollinearity:

a) high correlations between explanatory variables
b) instability of estimated coefficients (in response to changing specification)
c) the signs of coefficients do not conform to expectations
d) coefficients that should be statistically significant have large *se* and small *t*
e) the model R^2 is high, but there are no statistically significant coefficients

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

4. Variance Inflation Factors and Multicollinearity

A more rigorous method of detecting multicollinearity:

Recall that R_j^2 is coefficient of determination from regressing x_j on all other x.

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

If $VIF_j \geq 10$, then $R_j^2 \geq .90$.

Approach 1. run k regressions of each x_j on all $x_{\sim j}$

Approach 2. in R, use **vif(*model*)** in **car** package

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\sum(x_{ji}-\overline{x_j})^2}} \cdot \frac{1}{\sqrt{1-R_j^2}}$$

$$= \frac{\hat{\sigma}}{\sqrt{SST_j}} \cdot \sqrt{VIF_j}$$

for some useful diagnostic tools, see *http://www.statmethods.net/stats/rdiagnostics.html*

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

4. Variance Inflation Factors and Multicollinearity

Solutions to multicollinearity

1) drop a variable

2) use 'partialling out' method (depending on your theory… we discuss this in lecture 9)

3) create an index, or using principal components analysis

4) impose constraints on coefficients

5) expand the sample; Goldberger refers to this concern as "micronumerosity"

Note: see Tufte, *Data Analysis for Public Policy*, p. 148-155

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

5. Example: *Judicial Independence and Human Rights*

Dependent variable

Human Rights : measure based on adherence to rights enumerated in UN treaties
range is from 17 to 99, with a mean of 67 and s.d. of 24

Explanatory variables

Judicial Independence : measure based on tenure of judges and scope of judicial authority

Log GDP Per Capita : measures wealth

Democracy : measure based on ???

range is from 0 to 2, with mean of 1.07 and s.d. of 0.79

Note: *Judicial Independence* and *Democracy* correlate at $r = .26$

Sample is 63 countries; analysis found in Michael Bailey, *Real Econometrics*, p. 153–156, based on Anna Harvey (2011) "What's So Great About Independent Courts?" (ms)

Example is based on Rafael La Porta, F. Lopez-de-Silanes, C. Pop-Eleches, and A. Schleifer (2004) "Judicial Checks and Balances," *Journal of Political Economy* 112: 445–470

POLS 6481. Research Design and Quantitative Methods II
Lecture 8. Multiple Regression Properties; Multicollinearity
Readings: Wooldridge, *Introductory Econometrics 5e*, 3.3 + 3.4 + Appx 3A.5 + 4.2 + 5.2

5. Example: *Judicial Independence and Human Rights*

	Without democracy variable	With democracy variable
Judicial independence	11.37*	1.03
	(4.49)	(3.15)
	[$t = 2.53$]	[$t = 0.33$]
Log GDP per capita	9.77*	1.07
	(1.36)	(4.49)
	[$t = 7.20$]	[$t = 0.82$]
Democracy		24.93*
		(2.77)
		[$t = 9.01$]
Constant	−22.68	30.97*
	(12.57)	10.15
	[$t = 1.80$]	[$t = 3.05$]
N	63	63
$\hat{\sigma}$	17.6	11.5
R^2	0.47	0.78
$R^2_{Judicialind.}$		0.153
R^2_{LogGDP}		0.553
$R^2_{Democracy}$		0.552

Standard errors in parentheses.

* *indicates significance at* $p < 0.05$, *two-tailed.*

VIF	Std. Error Inflation
1.18	8.7%
2.24	49.6%
2.24	49.4%