# Linear Regression with Multiple Regressors (SW Ch. 6)

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w - Spring 2022

Brooklyn College
Apr 4, 2022

# Outline

1. Omitted variable bias
2. Multiple regression and OLS
3. Measures of fit
4. Sampling distribution of the OLS estimator with multiple regressors

# Omitted Variable Bias (SW Section 6.1) (1 of 5)

In the class size example, $\beta_1$ is the causal effect on test scores of a change in the *STR* by one student per teacher.

When $\beta_1$ is a causal effect, the first least squares assumption for causal inference must hold: $E(u|X) = 0$.

The error *u* arises because of factors, or variables, that influence *Y* but are not included in the regression function. There are always omitted variables!

If the omission of those variables results in $E(u|X) \neq 0$, then the OLS estimator will be biased.

# Omitted Variable Bias (SW Section 6.1) (2 of 5)

The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted variable "*Z*" must satisfy two conditions:

<u>The two conditions for omitted variable bias</u>

1. *Z* is a determinant of *Y* (i.e. *Z* is part of *u*); **and**

2. *Z* is correlated with the regressor *X* (*i.e.* corr($Z$,$X$) $\neq 0$)

***Both*** *conditions must hold for the omission of Z to result in omitted variable bias.*

# Omitted Variable Bias (SW Section 6.1) (3 of 5)

In the test score example:

1. English language ability (whether the student has English as a second language) plausibly affects standardized test scores: *Z* is a determinant of *Y*.

2. Immigrant communities tend to be less affluent and thus have smaller school budgets and higher *STR*: *Z* is correlated with *X*.

Accordingly, $\hat{\beta}_1$ is biased. What is the direction of this bias?

- *What does common sense suggest?*
- If common sense fails you, there is a formula…

# Omitted Variable Bias (SW Section 6.1) (4 of 5)

A formula for omitted variable bias: recall the equation,

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$.

Under Least Squares Assumption #1, $E[(X_i - \mu_X)u_i] = \text{cov}(X_i,u_i) = 0$.

But what if $E[(X_i - \mu_X)u_i] = \text{cov}(X_i,u_i) = \sigma_{Xu} \neq 0$?

# Omitted Variable Bias (SW Section 6.1) (5 of 5)

Let $\beta_1$ be the causal effect. Under LSA #2 and #3 (that is, even if LSA #1 does not hold),

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

$$= \left(\frac{\sigma_u}{\sigma_X}\right) \times \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu},$$

where $\rho_{Xu}$ = corr($X$,$u$). If assumption #1 is correct, then $\rho_{Xu}$ = 0, but if not we have….

# The omitted variable bias formula: (1 of 2)

$$\boxed{\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}}$$

- If an omitted variable $Z$ is ***both***:

1. a determinant of $Y$ (that is, it is contained in $u$); ***and***
2. correlated with $X$, then $\rho_{Xu} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased and is not consistent.

- For example, districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the effect of having many ESL students factor would result in overstating the class size effect. *Is this is actually going on in the CA data*?

# The omitted variable bias formula: (2 of 2)

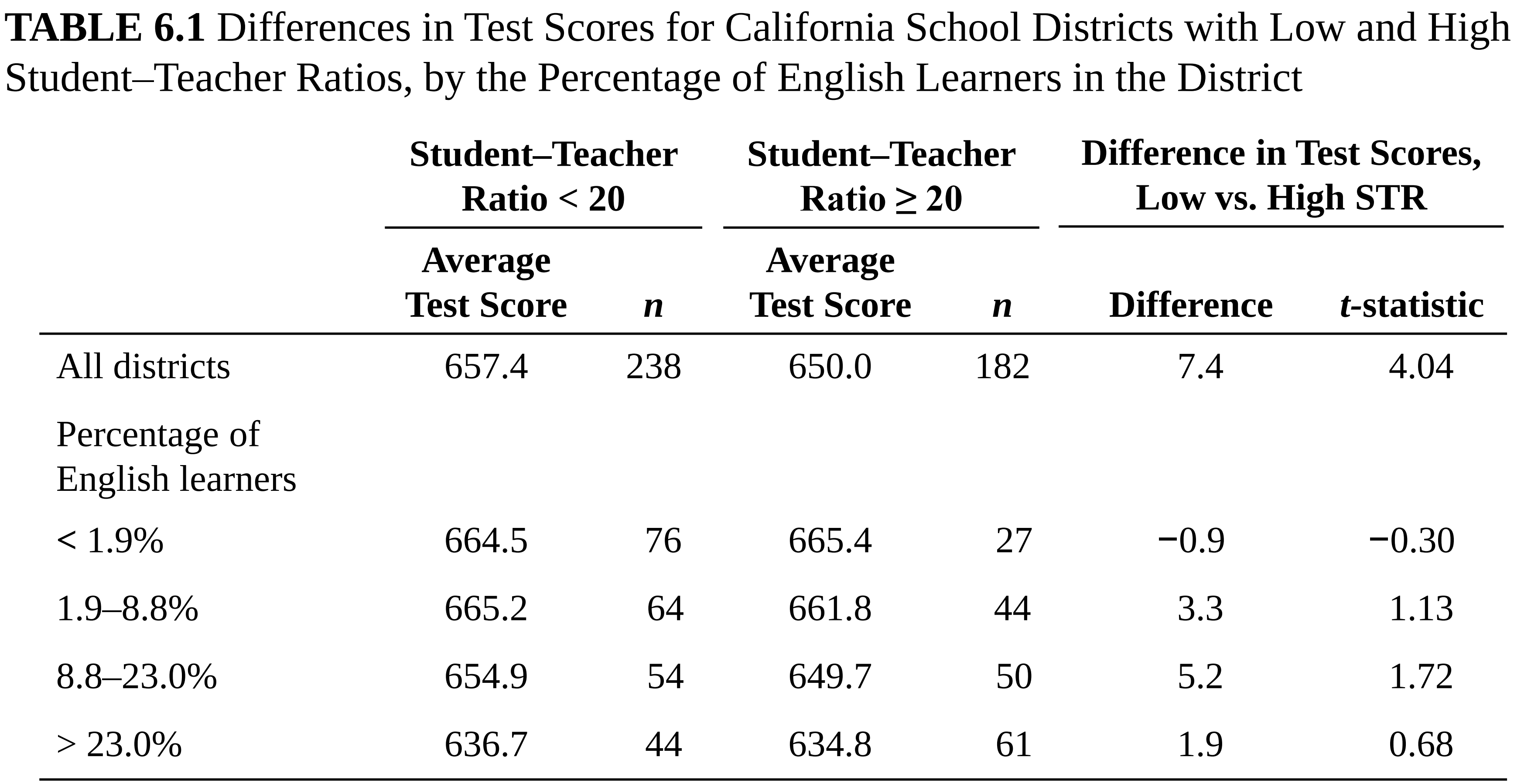

**TABLE 6.1** Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

| | **Student–Teacher Ratio < 20** | | **Student–Teacher Ratio ≥ 20** | | **Difference in Test Scores, Low vs. High STR** | |
|---|---|---|---|---|---|---|
| | **Average Test Score** | ***n*** | **Average Test Score** | ***n*** | **Difference** | ***t*-statistic** |
| All districts | 657.4 | 238 | 650.0 | 182 | 7.4 | 4.04 |
| Percentage of English learners | | | | | | |
| < 1.9% | 664.5 | 76 | 665.4 | 27 | −0.9 | −0.30 |
| 1.9–8.8% | 665.2 | 64 | 661.8 | 44 | 3.3 | 1.13 |
| 8.8–23.0% | 654.9 | 54 | 649.7 | 50 | 5.2 | 1.72 |
| > 23.0% | 636.7 | 44 | 634.8 | 61 | 1.9 | 0.68 |

- Districts with fewer English Learners have higher test scores
- Districts with lower percent *EL* (*PctEL*) have smaller classes
- Among districts with comparable *PctEL*, the effect of class size is small (recall overall "test score gap" = 7.4)

# *Return to omitted variable bias*

## Three ways to overcome omitted variable bias

1. Run a randomized controlled experiment in which treatment (*STR*) is randomly assigned: then *PctEL* is still a determinant of *TestScore*, but *PctEL* is uncorrelated with *STR*. (*This solution to OV bias is rarely feasible.*)

2. Adopt the “cross tabulation” approach, with finer gradations of *STR* and *PctEL* – within each group, all classes have the same *PctEL*, so we control for *PctEL* (*But soon you will run out of data, and what about other determinants like family income and parental education*?)

3. Use a regression in which the omitted variable (*PctEL*) is no longer omitted: include *PctEL* as an additional regressor in a multiple regression.

# The Population Multiple Regression Model (SW Section 6.2)

- Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1,\ldots,n$$

- *Y* is the *dependent variable*
- $X_1$, $X_2$ are the two *independent variables* (*regressors*)
- ($Y_i$, $X_{1i}$, $X_{2i}$) denote the $i^{\text{th}}$ observation on *Y*, $X_1$, and $X_2$.
- $\beta_0$ = unknown population intercept
- $\beta_1$ = effect on *Y* of a change in $X_1$, holding $X_2$ constant
- $\beta_2$ = effect on *Y* of a change in $X_2$, holding $X_1$ constant
- $u_i$ = the regression error (omitted factors)

# Interpretation of coefficients in multiple regression (1 of 2)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \ \ i = 1,\ldots,n$$

Consider the difference in the expected value of *Y* for two values of $X_1$ holding $X_2$ constant:

Population regression line when $X_1 = X_{1,0}$:

$$Y = \beta_0 + \beta_1 X_{1,0} + \beta_2 X_2$$

Population regression line when $X_1 = X_{1,0} + \Delta X_1$:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_{1,0} + \Delta X_1) + \beta_2 X_2$$

# Interpretation of coefficients in multiple regression (2 of 2)

***Before:*** $Y = \beta_0 + \beta_1(X_{1,0} + \Delta X_1) + \beta_2 X_2$

***After:*** $Y + \Delta Y = \beta_0 + \beta_1(X_{1,0} + \Delta X_1) + \beta_2 X_2$

***Difference:*** $\Delta Y = \beta_1 \Delta X_1$

***So:***

$$\boxed{\beta_1 = \frac{\Delta Y}{\Delta X_1}, \textbf{ holding } \boldsymbol{X_2} \textbf{ constant}}$$

$$\boxed{\beta_2 = \frac{\Delta Y}{\Delta X_2}, \textbf{ holding } \boldsymbol{X_1} \textbf{ constant}}$$

$$\boxed{\beta_0 = \text{predicted value of } Y \text{ when } X_1 = X_2 = 0.}$$

# The OLS Estimator in Multiple Regression (SW Section 6.3)

- With two regressors, the OLS estimator solves:

$$\min_{b_0, b_1, b_2} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of $Y_i$ and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **This yields the OLS estimators of $\beta_0$ , $\beta_1$ and $\beta_2$.**

# Example: the California test score data

Regression of *TestScore* against *STR*:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

Now include percent English Learners in the district (*PctEL*):

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65PctEL$$

- What happens to the coefficient on *STR*?

# Multiple regression in STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  223.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4264
                                                       Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472     -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318    -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224     78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65PctEL$$

*More on this printout later…*

# Measures of Fit for Multiple Regression (SW Section 6.4) (1 of 2)

Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$

$SER$ = std. deviation of $\hat{u}_i$ (with d.f. correction)

$RMSE$ = std. deviation of $\hat{u}_i$ (without d.f. correction)

$R^2$ = fraction of variance of $Y$ explained by $X$

$\bar{R}^2$ = "adjusted $R^2$" = $R^2$ with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$

# SER and RMSE

As in regression with a single regressor, the *SER* and the *RMSE* are measures of the spread of the *Y*s around the regression line:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

# $R^2$ and $\bar{R}^2$ (adjusted $R^2$) (1 of 2)

The $R^2$ is the fraction of the variance explained – same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

where $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2$, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$, $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

- The $R^2$ always increases when you add another regressor (*why?*) – a bit of a problem for a measure of “fit”

# $R^2$ and $\bar{R}^2$ (adjusted $R^2$) (2 of 2)

The $\bar{R}^2$ (the “adjusted $R^2$”) corrects this problem by “penalizing” you for including another regressor – the $\bar{R}^2$ does not necessarily increase when you add another regressor.

$$Adjusted\ R^2 : \bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$$

Note that $\bar{R}^2 < R^2$, however if $n$ is large the two will be very close.

# Measures of Fit for Multiple Regression (SW Section 6.4) (2 of 2)

Test score example:

(1) $\widehat{TestScore} = 698.9 - 2.28 \times STR,$

$$R^2 = .05, \;\; SER = 18.6$$

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL,$

$$R^2 = .426, \;\; \bar{R}^2 = .424, \;\; SER = 14.5$$

- *What – precisely – does this tell you about the fit of regression (2) compared with regression (1)?*
- *Why are the $R^2$ and the $\bar{R}^2$ so close in (2)?*

# The Least Squares Assumptions in Multiple Regression (SW Section 6.5)

Let $\beta_1$, $\beta_2$,…, $\beta_k$ be causal effects.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

1. The conditional distribution of $u$ given the $X$'s has mean zero, that is, $E(u_i|X_{1i} = x_1,\ldots, X_{ki} = x_k) = 0$.
2. $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are unlikely: $X_1,..., X_k$, and $Y$ have four moments: $E(X_{1i}^4) < \infty,..., E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$.
4. There is no perfect multicollinearity.

# Assumption #1: the conditional mean of *u* given the included *X*s is zero. (1 of 2)

$$E(u|X_1 = x_1,\ldots, X_k = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- Failure of this condition leads to omitted variable bias, specifically, if an omitted variable

  1. belongs in the equation (so is in *u*) ***and***
  2. is correlated with an included *X*

- then this condition fails and there is OV bias.
- The best solution, if possible, is to include the omitted variable in the regression.
- A second, related solution is to include a variable that controls for the omitted variable (discussed shortly)

# Assumption #1: the conditional mean of *u* given the included *X*s is zero. (2 of 2)

**Assumption #2: $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.**

This is satisfied automatically if the data are collected by simple random sampling.

**Assumption #3: large outliers are rare (finite fourth moments)**

This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).

# Assumption #4: There is no perfect multicollinearity

***Perfect multicollinearity*** is when one of the regressors is an exact linear function of the other regressors.

***Example***: Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
Regression with robust standard errors                 Number of obs =     420
                                                       F(  1,   418) =   19.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0512
                                                       Root MSE      =  18.581
------------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
    str |  (dropped)
  _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

## ***Perfect multicollinearity*** is when one of the regressors is an exact linear function of the other regressors.

- In the previous regression, $\beta_1$ is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant (???)
- We will return to perfect (and imperfect) multicollinearity shortly, with more examples…
- *With these least squares assumptions in hand, we now can derive the sampling distribution of* $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.

# The Sampling Distribution of the OLS Estimator (SW Section 6.6)

Under the four Least Squares Assumptions,

- The sampling distribution of $\hat{\beta}_1$ has mean $\beta_1$
- $\text{var}(\hat{\beta}_1)$ is inversely proportional to *n*.
- Other than its mean and variance, the exact (finite-*n*) distribution of $\hat{\beta}_1$ is very complicated; but for large *n*...
  - $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \overset{p}{\rightarrow} \beta_1$ (law of large numbers)
  - $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT)
  - These statements hold for $\hat{\beta}_1 ,\dots, \hat{\beta}_k$

*Conceptually, there is nothing new here!*

# Multicollinearity, Perfect and Imperfect (SW Section 6.7)

***Perfect multicollinearity*** is when one of the regressors is an exact linear function of the other regressors.

<u>Some more examples of perfect multicollinearity</u>

1. The example from before: you include *STR* twice,
2. Regress *TestScore* on a constant, *D*, and *B*, where: $D_i$ = 1 if *STR* ≤ 20, = 0 otherwise; $B_i$ = 1 if *STR* >20, = 0 otherwise, so $B_i = 1 – D_i$ and there is perfect multicollinearity.
3. Would there be perfect multicollinearity if the intercept (constant) were excluded from this regression? This example is a special case of…

# The dummy variable trap (1 of 2)

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other). If you include all these dummy variables *and* a constant, you will have perfect multicollinearity – this is sometimes called ***the dummy variable trap***.

- *Why is there perfect multicollinearity here*?
- *Solutions to the dummy variable trap*:
  1. Omit one of the groups (e.g. Senior), or
  2. Omit the intercept
- *What are the implications of (1) or (2) for the interpretation of the coefficients?*

# The dummy variable trap (2 of 2)

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or giving an error message or by “dropping” one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

# *Imperfect multicollinearity* (1 of 2)

Imperfect and perfect multicollinearity are quite different despite the similarity of the names.

***Imperfect multicollinearity*** occurs when two or more regressors are very highly correlated.

- Why the term “multicollinearity”? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are “co-linear” – but unless the correlation is exactly ±1, that collinearity is imperfect.

# *Imperfect multicollinearity* (2 of 2)

Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.

- The idea: the coefficient on $X_1$ is the effect of $X_1$ holding $X_2$ constant; but if $X_1$ and $X_2$ are highly correlated, there is very little variation in $X_1$ once $X_2$ is held constant – so the data don't contain much information about what happens when $X_1$ changes but $X_2$ doesn't.  If so, the variance of the OLS estimator of the coefficient on $X_1$ will be large.

- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.

- Imperfect multicollinearity manifests through high $R^2$ but high coefficient standard errors (i.e. statistically insignificant coefficients)

- The math?  See SW, App. 6.2