4.1 — Multivariate OLS Estimators

ECON 480 • Econometrics • Fall 2022

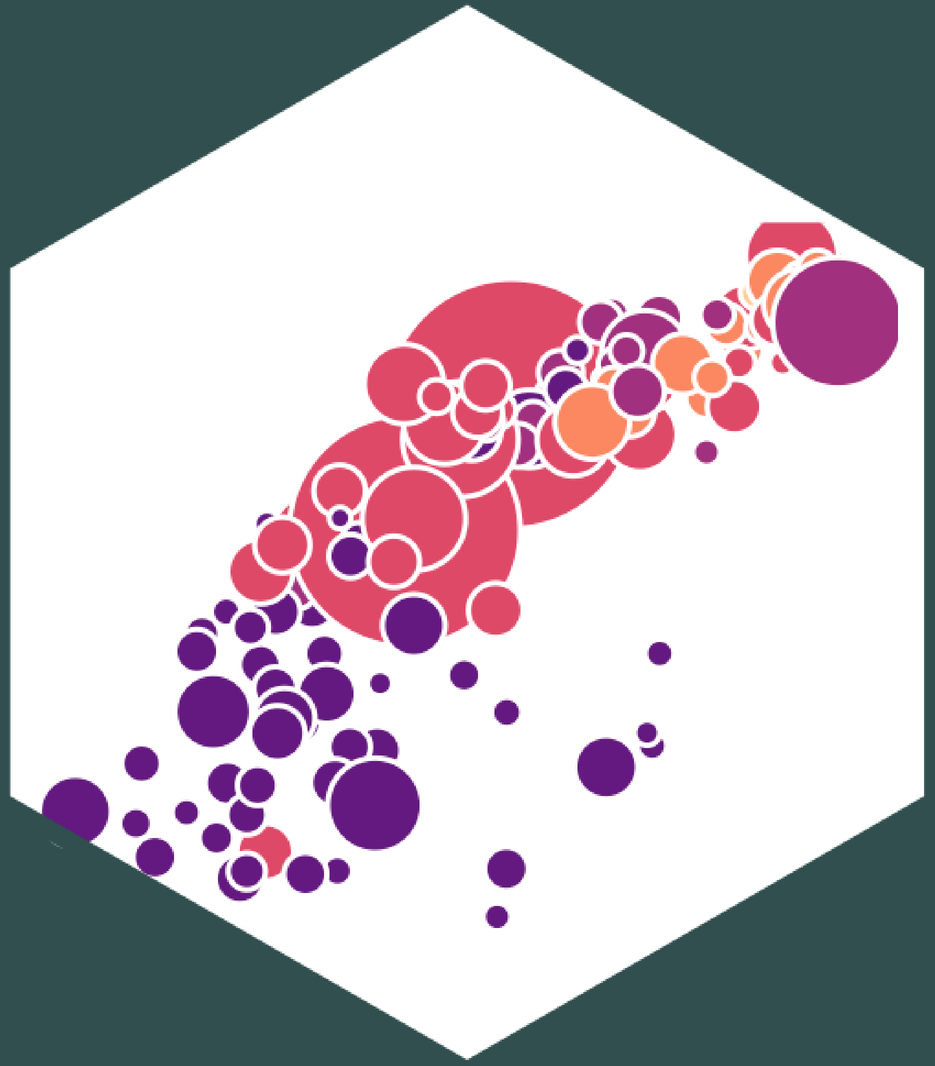
Dr. Ryan Safner

Associate Professor of Economics

safner@hood.edu

ryansafner/metricsF22

metricsF22.classes.ryansafner.com



Contents

The Multivariate OLS Estimators

The Expected Value of $\hat{\beta}_j$: Bias

Precision of $\hat{\beta}_j$

A Summary of Multivariate OLS Estimator Properties

(Updated) Measures of Fit

The Multivariate OLS Estimators

The Multivariate OLS Estimators

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

- The **ordinary least squares (OLS) estimators** of the unknown population parameters $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ solves:

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k} \sum_{i=1}^{n} \Bigg[\underbrace{Y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki})}_{\hat{Y}_i}}_{\hat{u}_i} \Bigg]^2$$

- Again, OLS estimators are chosen to **minimize** the **sum of squared residuals (SSR)**
  - i.e. sum of squared "distances" between actual values of Y_i and predicted values $\hat{Y}_i$

The Multivariate OLS Estimators: FYI

Math FYI

in linear algebra terms, a regression model with n observations of k independent variables:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}$$

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{Y}_{(n \times 1)}} = \underbrace{\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & \cdots & x_{k,n} \end{pmatrix}}_{\mathbf{X}_{(n \times k)}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta_{(k \times 1)}} + \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}}_{\mathbf{u}_{(n \times 1)}}$$

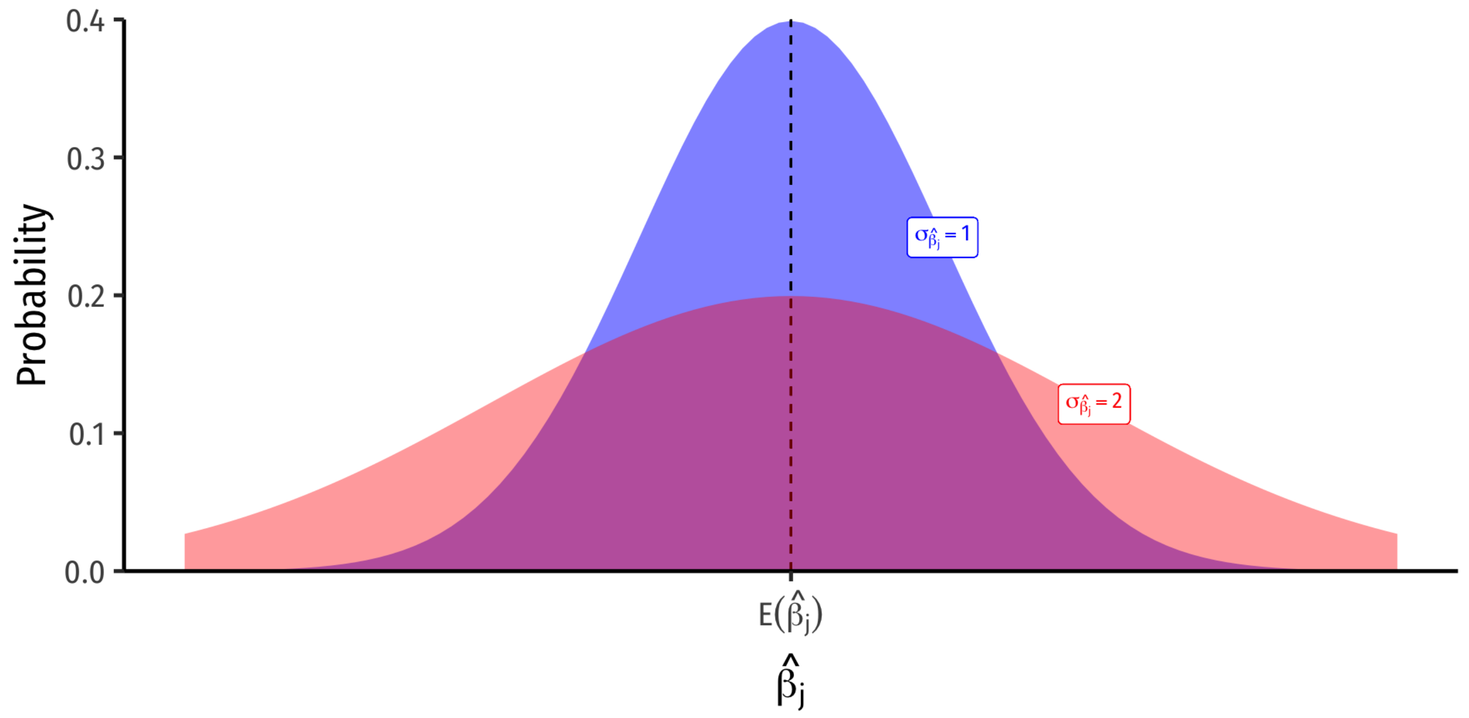
- The OLS estimator for β is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ 😱
- Appreciate that I am saving you from such sorrow 🤖

The Sampling Distribution of $\hat{\beta}_j$

- For *any* individual β_j, it has a sampling distribution:

$$\hat{\beta}_j \sim N\left(E[\hat{\beta}_j],\ se(\hat{\beta}_j)\right)$$

- We want to know its sampling distribution's:
  - **Center**: $E[\hat{\beta}_j]$; what is the *expected value* of our estimator?
  - **Spread**: $se(\hat{\beta}_j)$; how *precise* or *uncertain* is our estimator?

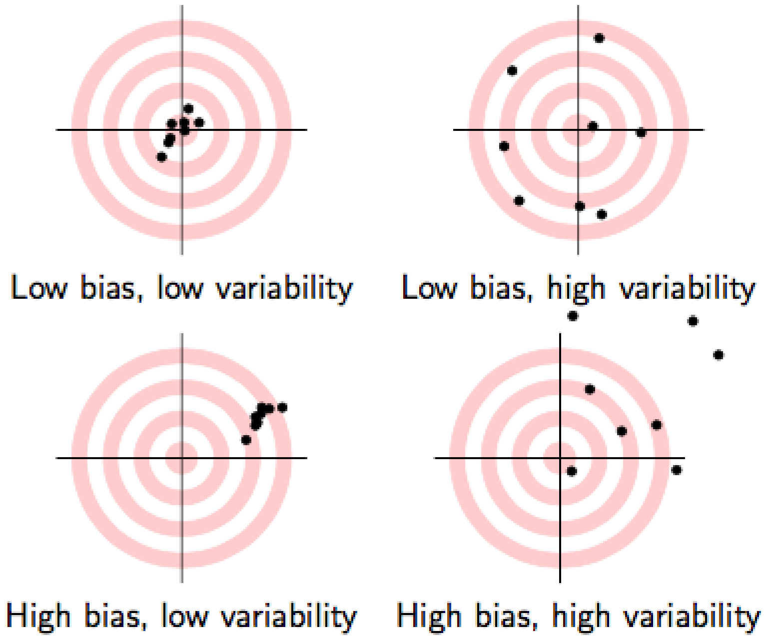


The Sampling Distribution of $\hat{\beta}_j$

- For *any* individual β_j, it has a sampling distribution:

$$\hat{\beta}_j \sim N\left(E[\hat{\beta}_j],\ se(\hat{\beta}_j)\right)$$

- We want to know its sampling distribution's:
  - **Center**: $E[\hat{\beta}_j]$; what is the *expected value* of our estimator?
  - **Spread**: $se(\hat{\beta}_j)$; how *precise* or *uncertain* is our estimator?



The Expected Value of $\hat{\beta}_j$: Bias

Exogeneity and Unbiasedness

- As before, $\mathbb{E}[\hat{\beta}_j] = \beta_j$ when X_j is **exogenous** (i.e. $cor(X_j, u) = 0$)

- We know the true $\mathbb{E}[\hat{\beta}_j] = \beta_j + \underbrace{cor(X_j, u)\frac{\sigma_u}{\sigma_{X_j}}}_{\text{O.V. Bias}}$

- If X_j is **endogenous** (i.e. $cor(X_j, u) \neq 0$), contains **omitted variable bias**

- Let's "see" an example of omitted variable bias and quantify it with our example

Measuring Omitted Variable Bias I

- Suppose the ***true* population model** of a relationship is:

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

- What happens when we run a regression and **omit** X_{2i}?
- Suppose we estimate the following **omitted regression** of just Y_i on X_{1i} (omitting X_{2i}):[1]

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \nu_i$$

1. Note: I am using α's and ν_i only to denote these are different estimates than the **true** model β's and u_i

Measuring Omitted Variable Bias II

- **Key Question:** are X_{1i} and X_{2i} correlated?

- Run an **auxiliary regression** of X_{2i} on X_{1i} to see:[1]

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + \tau_i$$

- If $\delta_1 = 0$, then X_{1i} and X_{2i} are *not* linearly related
- If $|\delta_1|$ is very big, then X_{1i} and X_{2i} are strongly linearly related

1. Note: I am using δ's and τ to differentiate estimates for this model.

Measuring Omitted Variable Bias III

- Now substitute our **auxiliary regression** between X_{2i} and X_{1i} into the ***true* model**:
  - We know $X_{2i} = \delta_0 + \delta_1 X_{1i} + \tau_i$

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

Measuring Omitted Variable Bias III

- Now substitute our **auxiliary regression** between X_{2i} and X_{1i} into the ***true* model**:
  - We know $X_{2i} = \delta_0 + \delta_1 X_{1i} + \tau_i$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 \big(\delta_0 + \delta_1 X_{1i} + \tau_i\big) + u_i \end{aligned}$$

Measuring Omitted Variable Bias III

- Now substitute our **auxiliary regression** between X_{2i} and X_{1i} into the ***true* model**:
  - We know $X_{2i} = \delta_0 + \delta_1 X_{1i} + \tau_i$

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 \big(\delta_0 + \delta_1 X_{1i} + \tau_i\big) + u_i \\
Y_i &= (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) X_{1i} + (\beta_2 \tau_i + u_i)
\end{aligned}
$$

Measuring Omitted Variable Bias III

- Now substitute our **auxiliary regression** between X_{2i} and X_{1i} into the ***true* model**:
  - We know $X_{2i} = \delta_0 + \delta_1 X_{1i} + \tau_i$

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 \big(\delta_0 + \delta_1 X_{1i} + \tau_i\big) + u_i \\
Y_i &= \underbrace{(\beta_0 + \beta_2 \delta_0)}_{\alpha_0} + \underbrace{(\beta_1 + \beta_2 \delta_1)}_{\alpha_1} X_{1i} + \underbrace{(\beta_2 \tau_i + u_i)}_{\nu_i}
\end{aligned}$$

- Now relabel each of the three terms as the OLS estimates (α's) and error (ν_i) from the **omitted regression**, so we again have:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \nu_i$$

- Crucially, this means that our OLS estimate for X_{1i} in the **omitted regression** is:

$$\alpha_1 = \beta_1 + \beta_2 \delta_1$$

Measuring Omitted Variable Bias IV

$$\alpha_1 = \beta_1 + \beta_2 \delta_1$$

- The **Omitted Regression** OLS estimate for X_1, (α_1) picks up *both*:

1. The true effect of X_1 on Y: β_1

2. The true effect of X_2 on Y: β_2...as pulled through the relationship between X_1 and X_2: δ_1

- Recall our conditions for omitted variable bias from some variable $\mathbf{Z_i}$:

1. $\mathbf{Z_i}$ **must be a determinant of** Y_i $\implies \beta_2 \neq 0$

2. $\mathbf{Z_i}$ **must be correlated with** X_i $\implies \delta_1 \neq 0$

- Otherwise, if Z_i does not fit these conditions, $\alpha_1 = \beta_1$ and the **omitted regression** is *unbiased*!

Measuring OVB in Our Class Size Example I

- The **“True” Regression** $(Y_i$ on X_{1i} and $X_{2i})$

$$\widehat{\text{Test Score}}_i = 686.03 - 1.10\ \text{STR}_i - 0.65\ \%\text{EL}_i$$

term <chr>	estimate <dbl>
(Intercept)	686.0322487
str	-1.1012959
el_pct	-0.6497768

3 rows | 1-2 of 5 columns

Measuring OVB in Our Class Size Example II

- The **"Omitted" Regression** (Y_i on just X_{1i})

$$\widehat{\text{Test Score}}_i = 698.93 - 2.28 \text{ STR}_i$$

term <chr>	estimate <dbl>
(Intercept)	698.932952
str	-2.279808

2 rows | 1-2 of 5 columns

Measuring OVB in Our Class Size Example III

- The **"Auxiliary" Regression** (X_{2i} on X_{1i})

$$\widehat{\%\text{EL}_i} = -19.85 + 1.81\ \text{STR}_i$$

term <chr>	estimate <dbl>
(Intercept)	-19.854055
str	1.813719

2 rows | 1-2 of 5 columns

Measuring OVB in Our Class Size Example IV

"True" Regression

$$\widehat{\text{Test Score}_i} = 686.03 - 1.10\ \text{STR}_i - 0.65\ \%\text{EL}$$

"Omitted" Regression

$$\widehat{\text{Test Score}_i} = 698.93 - 2.28\ \text{STR}_i$$

"Auxiliary" Regression

$$\widehat{\%\text{EL}_i} = -19.85 + 1.81\ \text{STR}_i$$

- Omitted Regression α_1 on STR is -2.28

Measuring OVB in Our Class Size Example IV

"True" Regression

$$\widehat{\text{Test Score}_i} = 686.03 - 1.10\ \text{STR}_i - 0.65\ \%\text{EL}$$

"Omitted" Regression

$$\widehat{\text{Test Score}_i} = 698.93 - 2.28\ \text{STR}_i$$

"Auxiliary" Regression

$$\widehat{\%\text{EL}_i} = -19.85 + 1.81\ \text{STR}_i$$

- Omitted Regression α_1 on STR is -2.28

$$\alpha_1 = \beta_1 + \beta_2\delta_1$$

- The true effect of STR on Test Score: -1.10

Measuring OVB in Our Class Size Example IV

"True" Regression

$$\widehat{\text{Test Score}_i} = 686.03 - 1.10\ \text{STR}_i - 0.65\ \%\text{EL}$$

"Omitted" Regression

$$\widehat{\text{Test Score}_i} = 698.93 - 2.28\ \text{STR}_i$$

"Auxiliary" Regression

$$\widehat{\%\text{EL}_i} = -19.85 + 1.81\ \text{STR}_i$$

- Omitted Regression α_1 on STR is -2.28

$$\alpha_1 = \beta_1 + \beta_2\delta_1$$

- The true effect of STR on Test Score: -1.10
- The true effect of %EL on Test Score: -0.65

Measuring OVB in Our Class Size Example IV

"True" Regression

$$\widehat{\text{Test Score}_i} = 686.03 - 1.10\ \text{STR}_i - 0.65\ \%\text{EL}$$

"Omitted" Regression

$$\widehat{\text{Test Score}_i} = 698.93 - 2.28\ \text{STR}_i$$

"Auxiliary" Regression

$$\widehat{\%\text{EL}_i} = -19.85 + 1.81\ \text{STR}_i$$

- Omitted Regression α_1 on STR is -2.28

$$\alpha_1 = \beta_1 + \beta_2\delta_1$$

- The true effect of STR on Test Score: -1.10
- The true effect of %EL on Test Score: -0.65
- The relationship between STR and %EL: 1.81

Measuring OVB in Our Class Size Example IV

"True" Regression

$$\widehat{\text{Test Score}_i} = 686.03 - 1.10\ \text{STR}_i - 0.65\ \%\text{EL}$$

"Omitted" Regression

$$\widehat{\text{Test Score}_i} = 698.93 - 2.28\ \text{STR}_i$$

"Auxiliary" Regression

$$\widehat{\%\text{EL}_i} = -19.85 + 1.81\ \text{STR}_i$$

- Omitted Regression α_1 on STR is -2.28

$$\alpha_1 = \beta_1 + \beta_2\delta_1$$

- The true effect of STR on Test Score: -1.10
- The true effect of %EL on Test Score: -0.65
- The relationship between STR and %EL: 1.81
- So, for the **omitted regression**:

$$-2.28 = -1.10 + (-0.65)(1.81)$$

Measuring OVB in Our Class Size Example IV

"True" Regression

$$\widehat{\text{Test Score}_i} = 686.03 - 1.10\ \text{STR}_i - 0.65\ \%\text{EL}$$

"Omitted" Regression

$$\widehat{\text{Test Score}_i} = 698.93 - 2.28\ \text{STR}_i$$

"Auxiliary" Regression

$$\widehat{\%\text{EL}_i} = -19.85 + 1.81\ \text{STR}_i$$

- Omitted Regression α_1 on STR is -2.28

$$\alpha_1 = \beta_1 + \beta_2 \delta_1$$

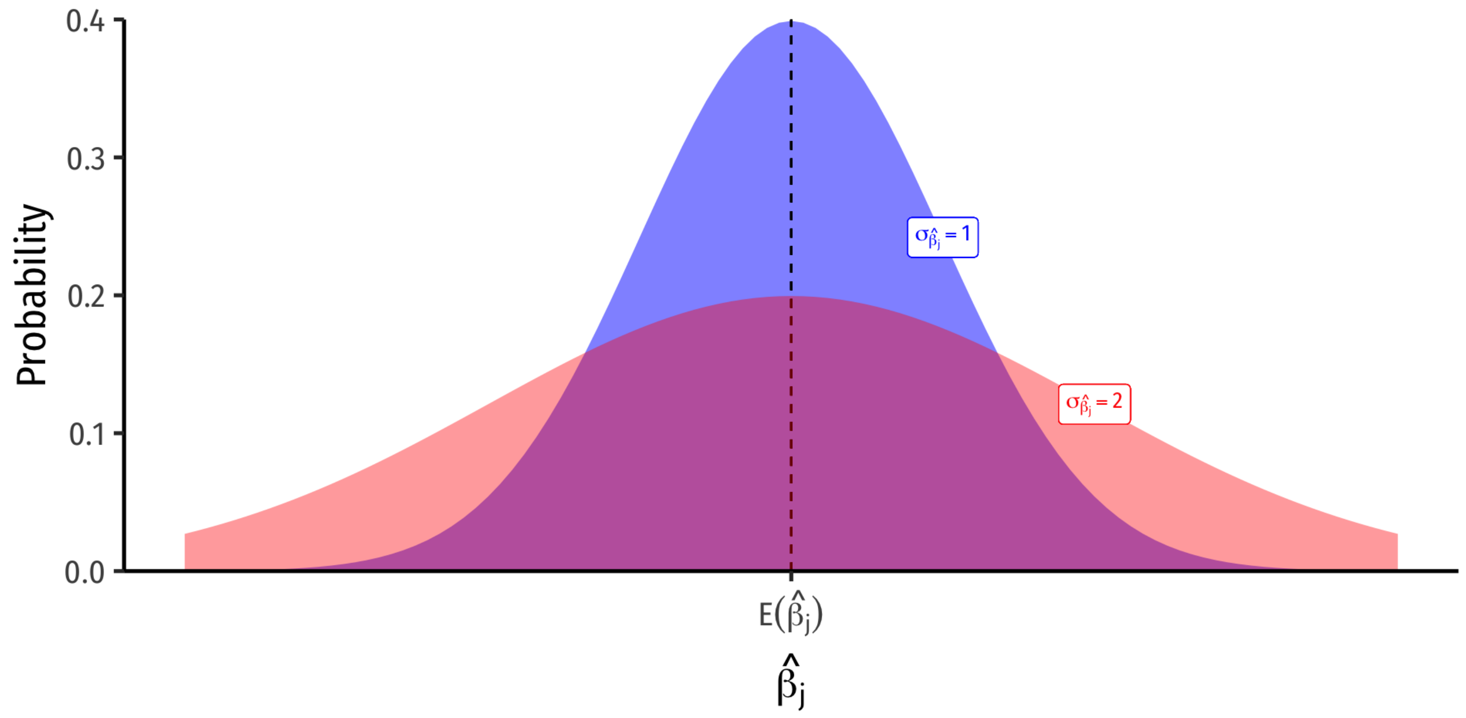
- The true effect of STR on Test Score: -1.10
- The true effect of %EL on Test Score: -0.65
- The relationship between STR and %EL: 1.81
- So, for the **omitted regression**:

$$-2.28 = -1.10 + \underbrace{(-0.65)(1.81)}_{O.V.Bias=\mathbf{-1.18}}$$

Precision of $\hat{\beta}_j$

Precision of $\hat{\beta}_j$ I

- $\sigma_{\hat{\beta}_j}$; how **precise** or **uncertain** are our estimates?
- **Variance $\sigma^2_{\hat{\beta}_j}$** or **standard error $\sigma_{\hat{\beta}_j}$**



Precision of $\hat{\beta}_j$ II

$$var(\hat{\beta}_j) = \underbrace{\frac{1}{1-R_j^2}}_{VIF} \times \frac{(SER)^2}{n \times var(X)}$$

$$se(\hat{\beta}_j) = \sqrt{var(\hat{\beta}_j)}$$

- Variation in $\hat{\beta}_j$ is affected by **four** things now[1]:

1. **Goodness of fit of the model (SER)**
   - Larger $SER \rightarrow$ larger $var(\hat{\beta}_j)$
2. **Sample size, *n***
   - Larger $n \rightarrow$ smaller $var(\hat{\beta}_j)$
3. **Variance of X**
   - Larger $var(X) \rightarrow$ smaller $var(\hat{\beta}_j)$
4. **Variance Inflation Factor** $\frac{1}{(1-R_j^2)}$
   - Larger VIF, larger $var(\hat{\beta}_j)$
   - **This is the only new effect**

1. See Class 2.5 for a reminder of variation with just one X variable.

VIF and Multicollinearity I

- Two *independent* (X) variables are **multicollinear**:

$$cor(X_j, X_l) \neq 0 \quad \forall j \neq l$$

- **Multicollinearity between X variables does *not bias* OLS estimates**
  - Remember, we pulled another variable out of u into the regression
  - If it were omitted, then it *would* cause omitted variable bias!

- **Multicollinearity does *increase the variance* of each OLS estimator** by

$$VIF = \frac{1}{(1 - R_j^2)}$$

VIF and Multicollinearity II

$$VIF = \frac{1}{(1-R_j^2)}$$

- R_j^2 is the R^2 from an **auxiliary regression** of X_j on all other regressors (X's)
  - i.e. proportion of $var(X_j)$ explained by other X's

VIF and Multicollinearity III

Example

Suppose we have a regression with three regressors ($k = 3$):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

- There will be three different R_j^2's, one for each regressor:

$$\begin{aligned} R_1^2 \text{ for } X_{1i} &= \gamma + \gamma X_{2i} + \gamma X_{3i} \\ R_2^2 \text{ for } X_{2i} &= \zeta_0 + \zeta_1 X_{1i} + \zeta_2 X_{3i} \\ R_3^2 \text{ for } X_{3i} &= \eta_0 + \eta_1 X_{1i} + \eta_2 X_{2i} \end{aligned}$$

VIF and Multicollinearity IV

$$VIF = \frac{1}{(1 - R_j^2)}$$

- R_j^2 is the R^2 from an **auxiliary regression** of X_j on all other regressors (X's)
  - i.e. proportion of $var(X_j)$ explained by other X's
- The R_j^2 tells us **how much *other* regressors explain regressor X_j**
- **Key Takeaway**: If other X variables explain X_j well (high R_J^2), it will be harder to tell how *cleanly* $X_j \rightarrow Y_i$, and so $var(\hat{\beta}_j)$ will be higher

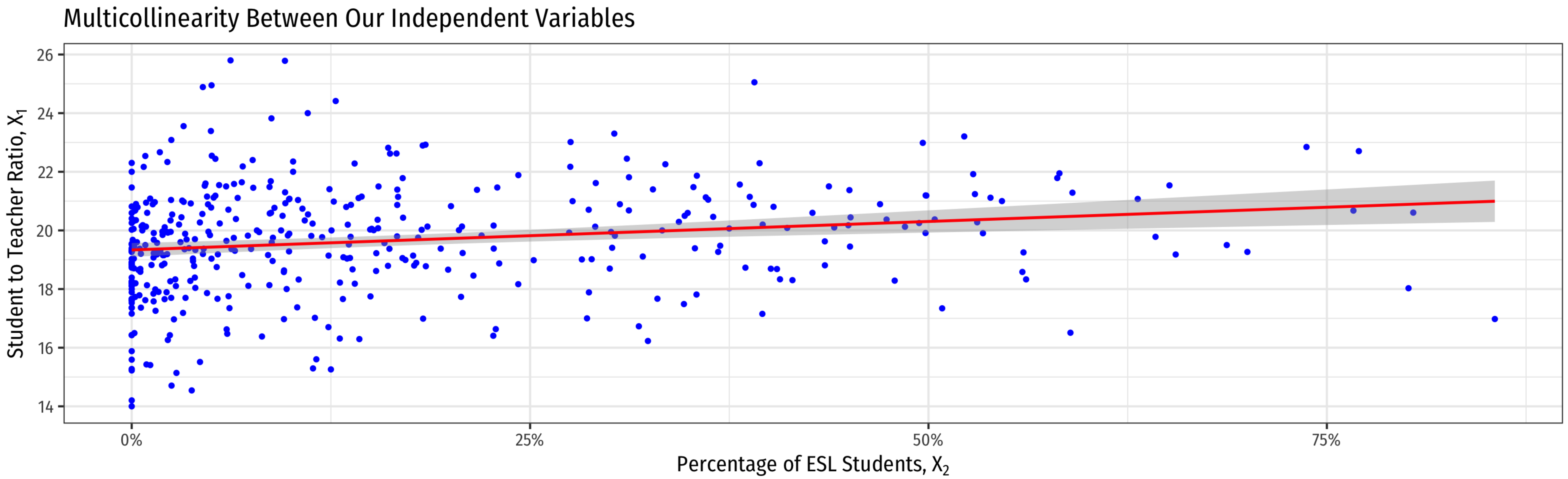
VIF and Multicollinearity V

- Common to calculate the **Variance Inflation Factor (VIF)** for each regressor:

$$VIF = \frac{1}{(1 - R_j^2)}$$

- VIF quantifies the factor (scalar) by which $var(\hat{\beta}_j)$ increases because of multicollinearity
  - e.g. VIF of 2, 3, etc. $\implies$ variance increases by 2x, 3x, etc.
- Baseline: $R_j^2 = 0 \implies$ *no* multicollinearity $\implies VIF = 1$ (no inflation)
- Larger $R_j^2 \implies$ larger VIF
  - Rule of thumb: $VIF > 10$ is problematic

VIF and Multicollinearity in Our Example I



- Higher $\%EL$ predicts higher STR
- Hard to get a precise marginal effect of STR holding $\%EL$ constant
  - Don't have much data on districts with *low* STR *and high* $\%EL$ (and vice versa)!

VIF and Multicollinearity in Our Example II

- Again, consider the correlation between the variables

```
ca_school %>%
  # Select only the three variables we want (there are many)
  select(str, testscr, el_pct) %>%
  # make a correlation table (all variables must be numeric)
  cor()
```

```
               str    testscr     el_pct
str      1.0000000 -0.2263628  0.1876424
testscr -0.2263628  1.0000000 -0.6441237
el_pct   0.1876424 -0.6441237  1.0000000
```

- $cor(STR, \%EL) = -0.644$

VIF and Multicollinearity in R I

```
# our multivariate regression
elreg <- lm(testscr ~ str + el_pct,
            data = ca_school)

# use the "car" package for VIF function
library("car")

elreg %>% vif()
```

```
     str   el_pct 
1.036495 1.036495 
```

- $var(\hat{\beta}_1)$ on `str` increases by **1.036** times (3.6%) due to multicollinearity with `el_pct`
- $var(\hat{\beta}_2)$ on `el_pct` increases by **1.036** times (3.6%) due to multicollinearity with `str`

VIF and Multicollinearity in R II

- Let's calculate VIF manually to see where it comes from:

```
# run auxiliary regression of x2 on x1
auxreg <- lm(el_pct ~ str,
             data = ca_school)

library(broom)
auxreg %>% tidy() # look at reg output
```

term <chr>	estimate <dbl>
(Intercept)	-19.854055
str	1.813719

2 rows | 1-2 of 5 columns

VIF and Multicollinearity in R III

```
auxreg %>% glance() # look at aux reg stats for R^2
```

r.squared <dbl>	adj.r.squared <dbl>
0.03520966	0.03290155

1 row | 1-2 of 12 columns

```
# extract our R-squared from aux regression (R_j^2)

aux_r_sq <- glance(auxreg) %>%
  pull(r.squared)

aux_r_sq # look at it
```

```
[1] 0.03520966
```

VIF and Multicollinearity in R IV

```
# calculate VIF manually

our_vif <- 1 / (1 - aux_r_sq) # VIF formula

our_vif
```

```
[1] 1.036495
```

- Again, multicollinearity between the two X variables inflates the variance on each by 1.036 times

Another Example: Expenditures/Student I

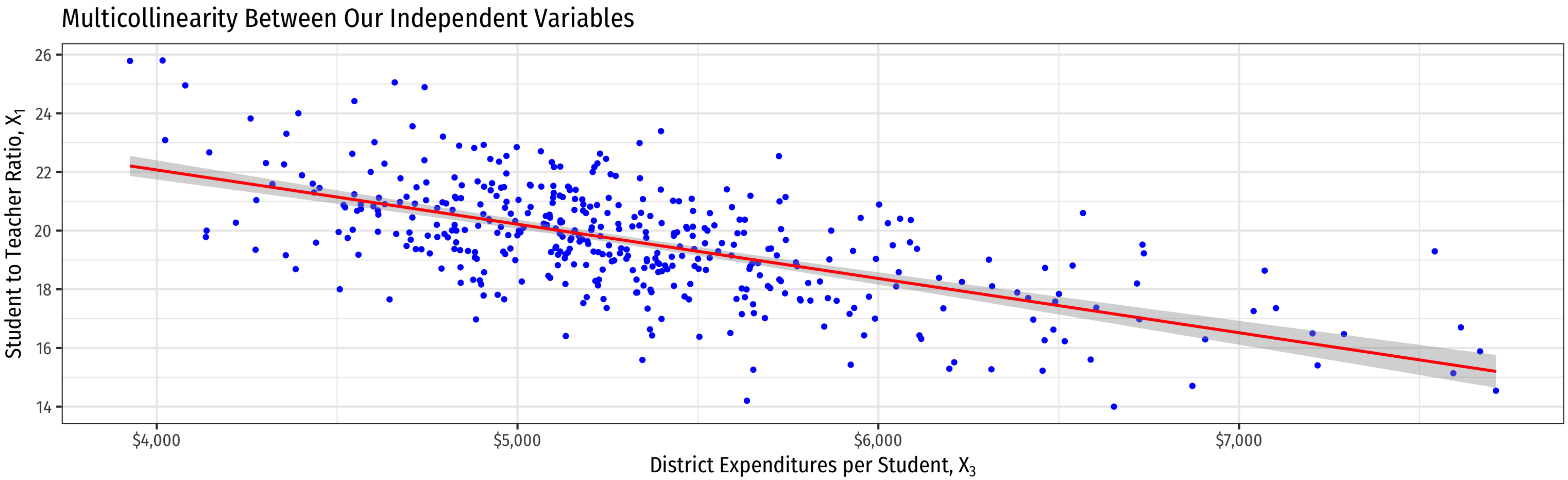
Example

What about district expenditures per student?

```
ca_school %>%
  select(testscr, str, el_pct, expn_stu) %>%
  cor()
```

```
           testscr        str      el_pct    expn_stu
testscr   1.0000000 -0.2263628 -0.64412374  0.19127277
str      -0.2263628  1.0000000  0.18764237 -0.61998215
el_pct   -0.6441237  0.1876424  1.00000000 -0.07139604
expn_stu  0.1912728 -0.6199821 -0.07139604  1.00000000
```

Another Example: Expenditures/Student II

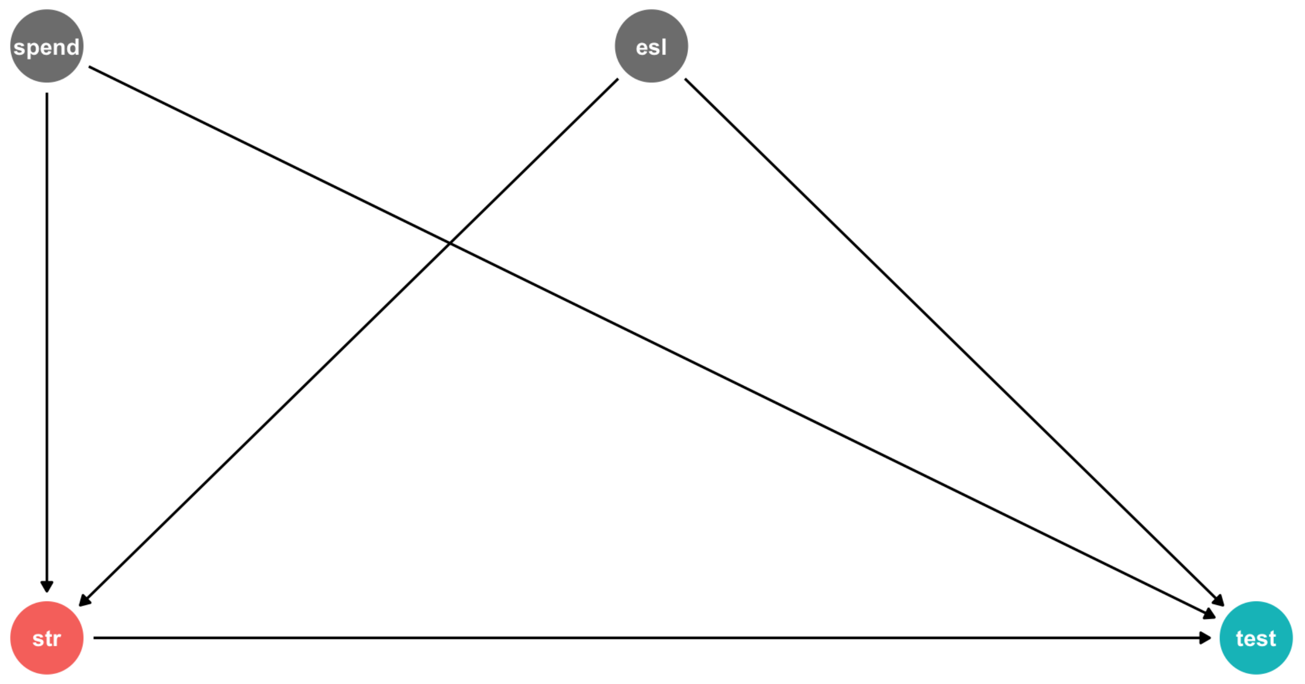


- Higher $spend$ predicts lower STR
- Hard to get a precise marginal effect of STR holding $spend$ constant
  - Don't have much data on districts with *high* STR *and high* $spend$ (and vice versa)!

Another Example: Expenditures/Student II

Would omitting Expenditures per student cause omitted variable bias?

1. $cor(Test, spend) \neq 0$
2. $cor(STR, spend) \neq 0$



Another Example: Expenditures/Student III

term <chr>	estimate <dbl>
(Intercept)	649.577947257
str	-0.286399240
el_pct	-0.656022660
expn_stu	0.003867902

4 rows | 1-2 of 5 columns

```r
vif(reg3)
```

```
     str   el_pct expn_stu 
1.680787 1.040031 1.629915 
```

- Including `expn_stu` reduces bias but increases variance of β_1 by 1.68x (68%)
  - and variance of β_2 by 1.04x (4%)

Multicollinearity Increases Variance

	Test Scores	Test Scores	Test Scores
Constant	698.93***	686.03***	649.58***
	(9.47)	(7.41)	(15.21)
Student Teacher Ratio	-2.28***	-1.10***	-0.29
	(0.48)	(0.38)	(0.48)
Percent ESL Students		-0.65***	-0.66***
		(0.04)	(0.04)
Spending per Student			0.00***
			(0.00)
n	420	420	420
R^2	0.05	0.43	0.44
SER	18.54	14.41	14.28

* p < 0.1, ** p < 0.05, *** p < 0.01

Perfect Multicollinearity

- ***Perfect* multicollinearity** is when a regressor is an exact linear function of (an)other regressor(s)

$$\widehat{Sales} = \hat{\beta}_0 + \hat{\beta}_1 \text{Temperature (C)} + \hat{\beta}_2 \text{Temperature (F)}$$

$$\text{Temperature (F)} = 32 + 1.8 * \text{Temperature (C)}$$

- $cor(\text{temperature (F), temperature (C)}) = 1$
- $R^2_j = 1 \rightarrow VIF = \frac{1}{1-1} \rightarrow var(\hat{\beta}_j) = 0$!
- **This is fatal for a regression**
  - A logical impossiblity, **always caused by human error**

Perfect Multicollinearity: Example

Example

$$\widehat{TestScore_i} = \hat{\beta}_0 + \hat{\beta}_1 STR_i + \hat{\beta}_2 \%EL + \hat{\beta}_3 \%EF$$

- $\%EL$: the percentage of students learning English
- $\%EF$: the percentage of students fluent in English
- $\%EF = 100 - \%EL$
- $|cor(\%EF, \%EL)| = 1$

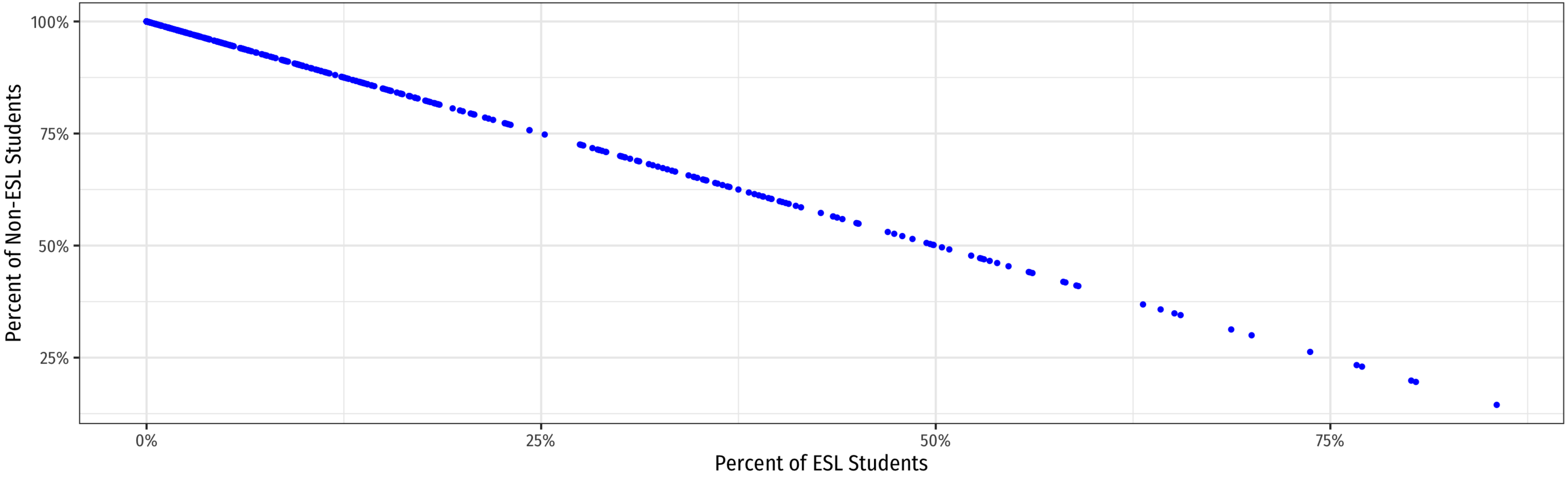
Perfect Multicollinearity: Example II

```r
# generate %EF variable from %EL
ca_school_ex <- ca_school %>%
  mutate(ef_pct = 100 - el_pct)

# get correlation between %EL and %EF
ca_school_ex %>%
  summarize(cor = cor(ef_pct, el_pct))
```

cor <dbl>
-1

1 row

Perfect Multicollinearity: Example III



Perfect Multicollinearity Example IV

```r
mcreg <- lm(testscr ~ str + el_pct + ef_pct,
            data = ca_school_ex)
summary(mcreg)
```

```
Call:
lm(formula = testscr ~ str + el_pct + ef_pct, data = ca_school_ex)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.845 -10.240  -0.308   9.815  43.461 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
str          -1.10130    0.38028  -2.896  0.00398 ** 
el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
ef_pct             NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264,	Adjusted R-squared:  0.4237 
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

```r
mcreg %>% tidy()
```

term <chr>
(Intercept)
str
el_pct
ef_pct

4 rows | 1-1 of 5 columns

- Note R *drops* one of the multicollinear regressors (`ef_pct`) if you include both 🤡

A Summary of Multivariate OLS Estimator Properties

A Summary of Multivariate OLS Estimator Properties

- $\hat{\beta}_j$ on X_j is biased only if there is an omitted variable (Z) such that:
  1. $cor(Y, Z) \neq 0$
  2. $cor(X_j, Z) \neq 0$
  - If Z is *included* and X_j is collinear with Z, this does *not* cause a bias
- $var[\hat{\beta}_j]$ and $se[\hat{\beta}_j]$ measure precision (or uncertainty) of estimate:

$$var[\hat{\beta}_j] = \frac{1}{(1 - R_j^2)} * \frac{SER^2}{n \times var[X_j]}$$

- VIF from multicollinearity: $\frac{1}{(1-R_j^2)}$
  - R_j^2 for auxiliary regression of X_j on all other X's
  - mutlicollinearity does not bias $\hat{\beta}_j$ but raises its variance
  - *perfect* multicollinearity if X's are linear function of others

(Updated) Measures of Fit

(Updated) Measures of Fit

- Again, how well does a linear model fit the data?
- How much variation in Y_i is "explained" by variation in the model $(\hat{Y_i})$?

$$
\begin{aligned}
Y_i &= \hat{Y_i} + \hat{u}_i \\
\hat{u}_i &= Y_i - \hat{Y_i}
\end{aligned}
$$

(Updated) Measures of Fit: SER

- Again, the **Standard errror of the regression (SER)** estimates the standard error of u

$$SER = \frac{SSR}{n - \mathbf{k} - 1}$$

- A measure of the spread of the observations around the regression line (in units of Y), the average "size" of the residual
- **Only new change:** divided by $n - k - 1$ due to use of $k + 1$ degrees of freedom to first estimate β_0 and then all of the other β's for the k number of regressors[1]

1. Again, because your textbook defines k as including the constant, the denominator would be $n-k$ instead of $n-k-1$

(Updated) Measures of Fit: R^2

$$\begin{aligned} R^2 &= \frac{SSM}{SST} \\ &= 1 - \frac{SSR}{SST} \\ &= (r_{X,Y})^2 \end{aligned}$$

- Again, R^2 is fraction of total variation in Y_i ("total sum of squares") that is explained by variation in predicted values ($\hat{Y}_i$), i.e. our model ("model sum of squares")

$$R^2 = \frac{var(\hat{Y})}{var(Y)}$$

Visualizing R^2

- **Total Variation in Y**: Areas **A** + D + E + G

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- **Variation in Y explained by X1 and X2**: Areas D + E + G

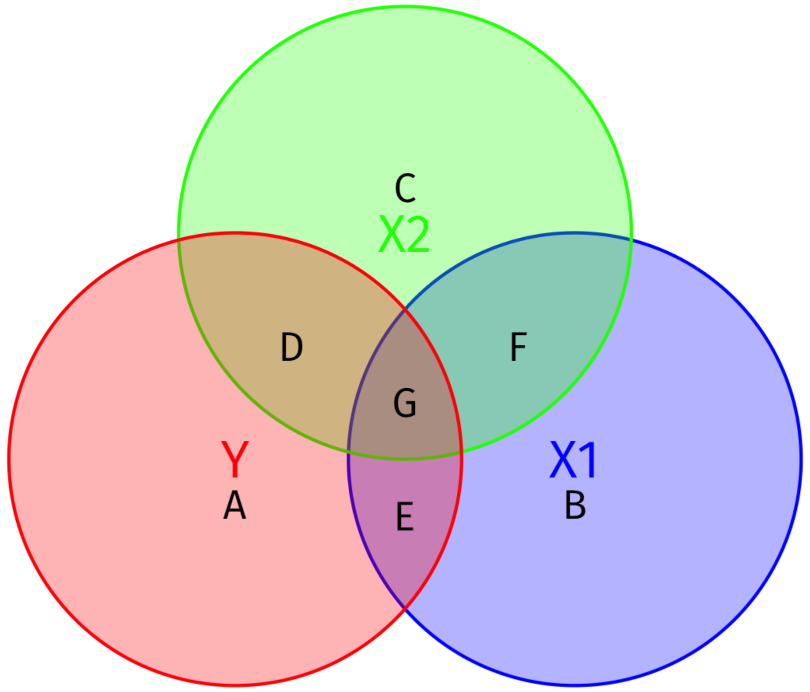
$$SSM = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

- **Unexplained variation in Y**: **Area A**

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

Compare with one X variable

$$R^2 = \frac{SSM}{SST} = \frac{D+E+G}{A+D+E+G}$$

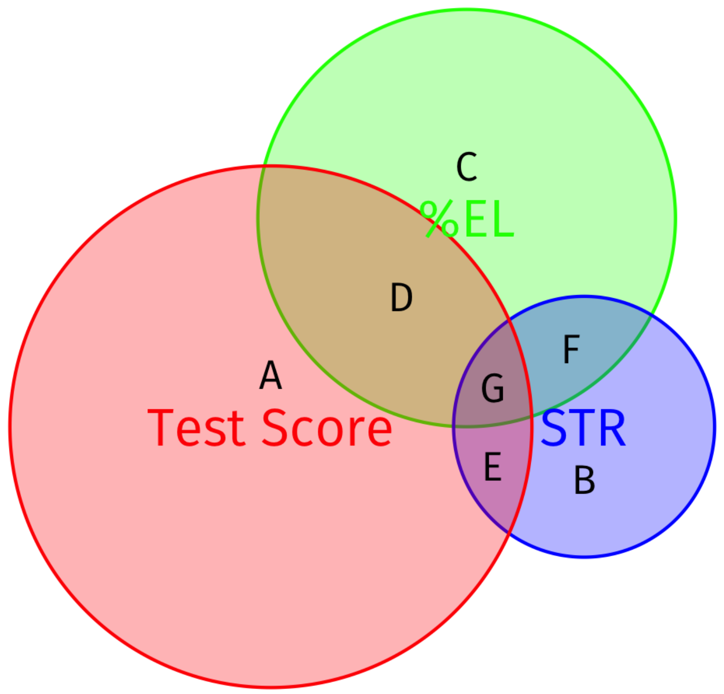


Visualizing R^2

```r
# make a function to calc. sum of sq. devs
sum_sq <- function(x){sum((x - mean(x))^2)}

# find total sum of squares
SST <- elreg %>%
  augment() %>%
  summarize(SST = sum_sq(testscr))

# find explained sum of squares
SSM <- elreg %>%
  augment() %>%
  summarize(SSM = sum_sq(.fitted))

# look at them and divide to get R^2
tribble(
  ~SSM, ~SST, ~R_sq,
  SSM, SST, SSM/SST
  ) %>%
  knitr::kable()
```

SSM	SST	R_sq
64864.3	152109.6	0.4264314

$$R^2 = \frac{SSM}{SST} = \frac{D+E+G}{A+D+E+G}$$



(Updated) Measures of Fit: Adjusted $\bar{R}^2$

- Problem: R^2 **mechanically** increases *every* time a new variable is added (it reduces SSR!)
  - Think in the diagram: more area of Y covered by more X variables!
- This does **not** mean adding a variable *improves the fit of the model* per se, R^2 gets **inflated**
- We correct for this effect with the **adjusted** $\bar{R}^2$ which penalizes adding new variables:

$$\bar{R}^2 = 1 - \underbrace{\frac{n-1}{n-k-1}}_{penalty} \times \frac{SSR}{SST}$$

- In the end, recall R^2 **was never that useful**[1], so don't worry about the formula
  - Large sample sizes (n) make R^2 and $\bar{R}^2$ very close

$\bar{R}^2$ In R

```
1  summary(elreg)
```

```
Call:
lm(formula = testscr ~ str + el_pct, data = ca_school)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.845 -10.240  -0.308   9.815  43.461 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 686.03225    7.41131  92.566  < 2e-16 ***
str          -1.10130    0.38028  -2.896  0.00398 ** 
el_pct       -0.64978    0.03934 -16.516  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.46 on 417 degrees of freedom
Multiple R-squared:  0.4264,    Adjusted R-squared:  0.4237 
F-statistic:   155 on 2 and 417 DF,  p-value: < 2.2e-16
```

- Base R^2 (`R` calls it "`Multiple R-squared`") went up
- `Adjusted R-squared` ($\bar{R}^2$) went down

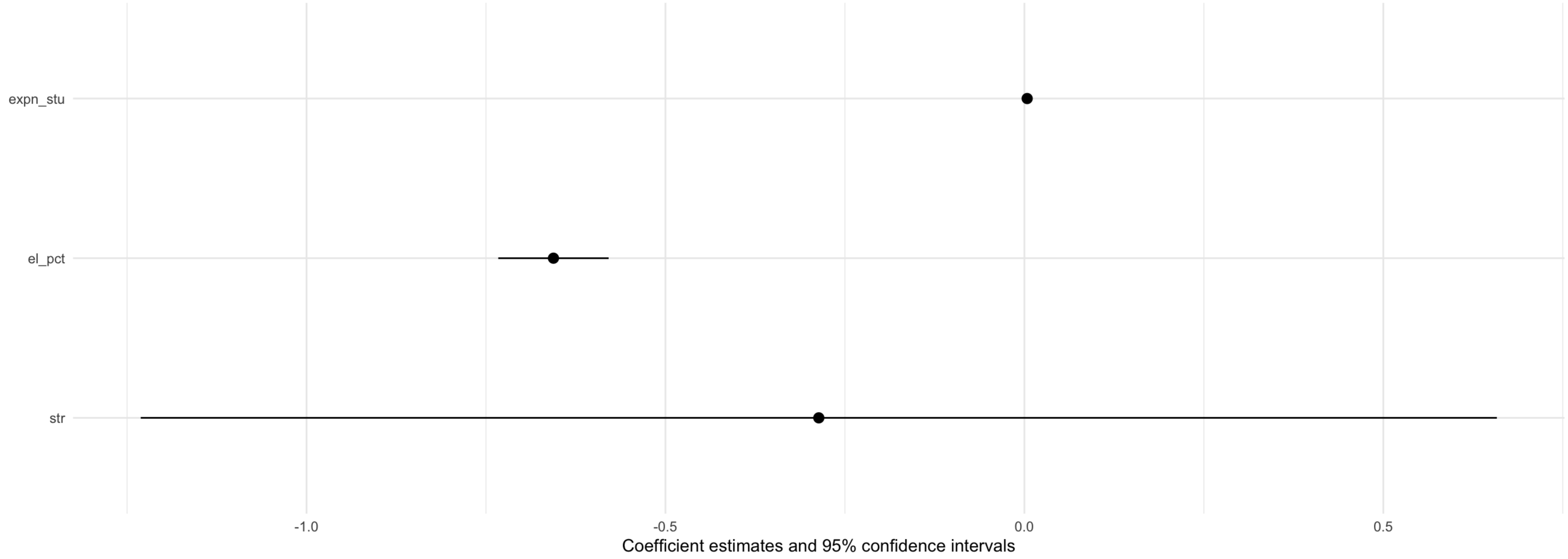
```
1  glance(elreg)
```

r.squared <dbl>
0.4264314

1 row | 1-1 of 12 columns

Coefficient Plots (with `modelsummary`)

Plot | Code



Regression Table (with `modelsummary`)

Output | Code

	Simple Model	MV Model 1	MV Model 2
Constant	698.93***	686.03***	649.58***
	(9.47)	(7.41)	(15.21)
STR	-2.28***	-1.10***	-0.29
	(0.48)	(0.38)	(0.48)
% ESL Students		-0.65***	-0.66***
		(0.04)	(0.04)
Spending per Student			0.00***
			(0.00)
N	420	420	420
Adj. R^2	0.05	0.42	0.43
SER	18.54	14.41	14.28

* p < 0.1, ** p < 0.05, *** p < 0.01