Lecture 11: Multivariate OLS Estimators

ECON 480 - Econometrics - Fall 2018

Ryan Safner

October 29, 2018

The Multivariate OLS Estimators

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

(Updated) Measures of Fit

The Multivariate OLS Estimators

- By analogy, we still focus on the ordinary least squares (OLS) estimators of the unknown population parameters β_0, β_1, β_2, $\cdots$, β_k which solves:

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

The OLS estimators minimize SSE, i.e. the sum of the squared distances between the actual values of Y_i and the predicted values $\hat{Y}_i$

Math FYI: Advanced Econometrics

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

$$\mathsf{Y} = \mathsf{X}\beta + \epsilon$$

Math FYI: Advanced Econometrics

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

$$Y = X\beta + \epsilon$$

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y_{(n\times 1)}} = \underbrace{\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & \cdots & x_{k,n} \end{pmatrix}}_{X_{(n\times k)}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta_{(k\times 1)}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\epsilon_{(n\times 1)}}$$

Math FYI: Advanced Econometrics

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

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

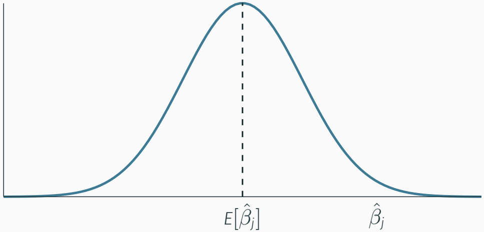
$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{Y}_{(n \times 1)}} = \underbrace{\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & \cdots & x_{k,n} \end{pmatrix}}_{\mathbf{X}_{(n \times k)}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta_{(k \times 1)}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\epsilon_{(n \times 1)}}$$

- The OLS estimator for β is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

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

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

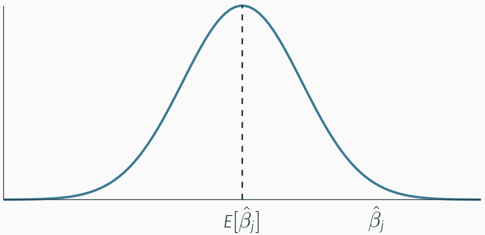
- We want to know:[1]



[1] I am using β_j to mean any one the k number of β's (associated any one X_j of the k X variables in our model. We've already used i to refer to any individual observation, and k to refer to the final variable, so I'm using j.

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

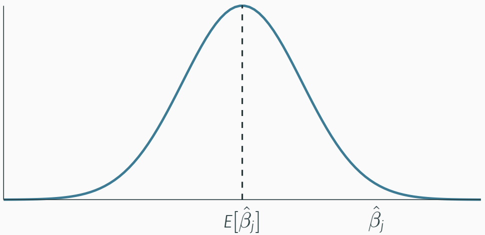
- We want to know:[1]
  - $E[\hat{\beta}_j]$; what is the expected value of our estimator?



[1] I am using β_j to mean any one the k number of β's (associated any one X_j of the k X variables in our model. We've already used i to refer to any individual observation, and k to refer to the final variable, so I'm using j.

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

- We want to know:[1]
  - $E[\hat{\beta}_j]$; what is the expected value of our estimator?
  - $SE(\hat{\beta}_j)$; how precise is our estimator?



[1] I am using β_j to mean any one the k number of β's (associated any one X_j of the k X variables in our model. We've already used i to refer to any individual observation, and k to refer to the final variable, so I'm using j.

- As before, we said that $E[\hat{\beta}_j] = \beta_j$ when X_j is **exogenous** (i.e. $corr(X_j, \epsilon) = 0$)

- As before, we said that $E[\hat{\beta}_j] = \beta_j$ when X_j is **exogenous** (i.e. $corr(X_j, \epsilon) = 0$)
- We know the true $E[\hat{\beta}_j] = \beta_j + \underbrace{corr(X_j, \epsilon)\frac{\sigma_\epsilon}{\sigma_{X_j}}}_{\text{O.V. Bias}}$

- As before, we said that $E[\hat{\beta}_j] = \beta_j$ when X_j is **exogenous** (i.e. $corr(X_j, \epsilon) = 0$)
- We know the true $E[\hat{\beta}_j] = \beta_j + \underbrace{corr(X_j, \epsilon)\frac{\sigma_\epsilon}{\sigma_{X_j}}}_{\text{O.V. Bias}}$
- We can now try to quantify the omitted variable bias

- Suppose the true population model of a relationship is:

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

[2] Note: I am using α's and ν_i only to denote these are different estimates than the true model β's and ϵ_i

- Suppose the true population model of a relationship is:

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

- What happens when we **omit** X_{2i}?

[2] Note: I am using α's and ν_i only to denote these are different estimates than the true model β's and ϵ_i

- Suppose the true population model of a relationship is:

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

- What happens when we **omit** X_{2i}?
- We estimate the following omitted regression leaving out X_{2i}:[2]

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

[2] Note: I am using α's and ν_i only to denote these are different estimates than the true model β's and ϵ_i

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

[3] Again, I'm using δ's and τ to differentiate estimates for this model

- **Key Question**: are X_{1i} and X_{2i} correlated?
- Run an auxiliary regression of X_{2i} on X_{1i}:[3]

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

[3] Again, I'm using δ's and τ to differentiate estimates for this model

- **Key Question**: are X_{1i} and X_{2i} correlated?
- Run an auxiliary regression of X_{2i} on X_{1i}:[3]

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

- If $\delta_1 = 0$, then X_{1i} and X_{2i} are not linearly related

[3] Again, I'm using δ's and τ to differentiate estimates for this model

- **Key Question**: are X_{1i} and X_{2i} correlated?
- Run an auxiliary regression of X_{2i} on X_{1i}:[3]

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

- 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

[3] Again, I'm using δ's and τ to differentiate estimates for this model

- Now substitute our auxiliary regression between X_{2i} and X_{1i} into the true model:

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

- Now substitute our auxiliary regression between X_{2i} and X_{1i} into the true model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \big(\delta_0 + \delta_1 X_{1i} + \tau_i\big) + \epsilon_i$$

Measuring Omitted Variable Bias III

- Now substitute our auxiliary regression between X_{2i} and X_{1i} into the true model:

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

HOOD COLLEGE

- Now substitute our auxiliary regression between X_{2i} and X_{1i} into the true model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \big(\delta_0 + \delta_1 X_{1i} + \tau_i\big) + \epsilon_i$$
$$Y_i = (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) X_{1i} + (\beta_2 \tau_i + \epsilon_i)$$

- Relabel each of the three terms as the OLS estimates (α's) from the omitted regression

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

- Now substitute our auxiliary regression between X_{2i} and X_{1i} into the true model:

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

- Relabel each of the three terms as the OLS estimates (α's) from the omitted regression

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

- Now substitute our auxiliary regression between X_{2i} and X_{1i} into the true model:

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

- Relabel each of the three terms as the OLS estimates (α's) from the omitted regression

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

- Now substitute our auxiliary regression between X_{2i} and X_{1i} into the true model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \left(\delta_0 + \delta_1 X_{1i} + \tau_i\right) + \epsilon_i$$
$$Y_i = (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) X_{1i} + (\beta_2 \tau_i + \epsilon_i)$$

- Relabel each of the three terms as the OLS estimates (α's) from the omitted regression

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

- This means that

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

- This means that:

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

- This means that:

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

- The Omitted Regression OLS estimate for X_{1i} picks up both:

- This means that:

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

- The Omitted Regression OLS estimate for X_{1i} picks up both:
    1. The true effect of X_{1i} on Y_i (β_1)

- This means that:

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

- The Omitted Regression OLS estimate for X_{1i} picks up both:
  1. The true effect of X_{1i} on Y_i (β_1)
  2. The effect of X_{2i} on Y_i (β_2)

- This means that:

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

- The Omitted Regression OLS estimate for X_{1i} picks up both:

1. The true effect of X_{1i} on Y_i (β_1)
2. The effect of X_{2i} on Y_i (β_2)
    - as pulled through the relationship between X_{2i} and X_{1i} (δ_1)

Measuring Omitted Variable Bias IV

- This means that:

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

- The Omitted Regression OLS estimate for X_{1i} picks up both:
  1. The true effect of X_{1i} on Y_i (β_1)
  2. The effect of X_{2i} on Y_i (β_2)
     - as pulled through the relationship between X_{2i} and X_{1i} (δ_1)
- Again, recall our conditions for omitted variable bias:

- This means that:

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

- The Omitted Regression OLS estimate for X_{1i} picks up both:
  1. The true effect of X_{1i} on Y_i (β_1)
  2. The effect of X_{2i} on Y_i (β_2)
     - as pulled through the relationship between X_{2i} and X_{1i} (δ_1)
- Again, recall our conditions for omitted variable bias:
  1. Z_i must be a determinant of Y_i: ($\beta_2 \neq 0$)

- This means that:

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

- The Omitted Regression OLS estimate for X_{1i} picks up both:
  1. The true effect of X_{1i} on Y_i (β_1)
  2. The effect of X_{2i} on Y_i (β_2)
     - as pulled through the relationship between X_{2i} and X_{1i} (δ_1)
- Again, recall our conditions for omitted variable bias:
  1. Z_i must be a determinant of Y_i: ($\beta_2 \neq 0$)
  2. Z_i is correlated with X_i: ($\delta_1 \neq 0$)

Measuring Omitted Variable Bias: Example

```r
true<-lm(testscr~str+el_pct, data=CASchool)
summary(true)
```

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

The ”True” Regression (Y on X_1 and X_2)

$$\widehat{\text{Test Score}} = \underset{(7.41)}{686.03} - \underset{(0.38)}{1.10}\, STR - \underset{(0.04)}{0.65}\, \%EL$$

Measuring Omitted Variable Bias: Example II

```r
omitted<-lm(testscr~str, data=CASchool)
summary(omitted)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

The "Omitted" Regression (Y on X_1)

$$\widehat{\text{Test Score}} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\,\text{STR}$$

Measuring Omitted Variable Bias: Example III

```r
auxiliary<-lm(el_pct~str, data=CASchool)
summary(auxiliary)
```

```
##
## Call:
## lm(formula = el_pct ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.823 -13.006  -6.849   7.834  74.601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.8541     9.1626  -2.167  0.03081 *
## str           1.8137     0.4644   3.906  0.00011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.98 on 418 degrees of freedom
## Multiple R-squared:  0.03521,    Adjusted R-squared:  0.0329
## F-statistic: 15.25 on 1 and 418 DF,  p-value: 0.0001095
```

The ”Auxiliary” Regression (X_2 on X_1)

$$\widehat{\%\ EL} = -19.85 + 1.81\ STR$$
$$(9.16) \qquad (0.46)$$

Measuring Omitted Variable Bias: Example IV

True Model:

$$\widehat{\text{Test Score}} = \underset{(7.41)}{686.03} - \underset{(0.38)}{1.10}\,\text{STR} - \underset{(0.04)}{0.65}\,\%\ \text{EL}$$

Omitted Regression:

$$\widehat{\text{Test Score}} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\,\text{STR}$$

Auxiliary Regression:

$$\widehat{\%\ \text{EL}} = \underset{(9.16)}{-19.85} + \underset{(0.46)}{1.81}\,\text{STR}$$

- Omitted Regression estimate for α_1 on STR is -2.28.

True Model:

$$\widehat{\text{Test Score}} = 686.03 - 1.10\,STR - 0.65\,\%\ EL$$
$$(7.41) \quad (0.38) \quad (0.04)$$

Omitted Regression:

$$\widehat{\text{Test Score}} = 698.93 - 2.28\,STR$$
$$(9.47) \quad (0.48)$$

Auxiliary Regression:

$$\widehat{\%\ EL} = -19.85 + 1.81\,STR$$
$$(9.16) \quad (0.46)$$

- Omitted Regression estimate for α_1 on STR is -2.28.

$$\alpha_1 = \beta_1 + \underbrace{\beta_2 \delta_1}_{bias}$$

True Model:

$$\widehat{\text{Test Score}} = \underset{(7.41)}{686.03} - \underset{(0.38)}{1.10}\,STR - \underset{(0.04)}{0.65}\,\%\ EL$$

Omitted Regression:

$$\widehat{\text{Test Score}} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\,STR$$

Auxiliary Regression:

$$\widehat{\%\ EL} = \underset{(9.16)}{-19.85} + \underset{(0.46)}{1.81}\,STR$$

- Omitted Regression estimate for α_1 on STR is -2.28.

$$\alpha_1 = \beta_1 + \underbrace{\beta_2\delta_1}_{bias}$$

- $\beta_1 = -1.10$ (True effect of STR on Test Score)

True Model:

$$\widehat{\text{Test Score}} = \underset{(7.41)}{686.03} - \underset{(0.38)}{1.10}\,STR - \underset{(0.04)}{0.65}\,\%\ EL$$

Omitted Regression:

$$\widehat{\text{Test Score}} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\,STR$$

Auxiliary Regression:

$$\widehat{\%\ EL} = \underset{(9.16)}{-19.85} + \underset{(0.46)}{1.81}\,STR$$

- Omitted Regression estimate for α_1 on STR is -2.28.

$$\alpha_1 = \beta_1 + \underbrace{\beta_2 \delta_1}_{bias}$$

- $\beta_1 = -1.10$ (True effect of STR on Test Score)
- $\beta_2 = -0.65$ (True effect of % EL on Test Score)

True Model:

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

(7.41) (0.38) (0.04)

Omitted Regression:

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

(9.47) (0.48)

Auxiliary Regression:

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

(9.16) (0.46)

- Omitted Regression estimate for α_1 on STR is -2.28.

$$\alpha_1 = \beta_1 + \underbrace{\beta_2\delta_1}_{bias}$$

- $\beta_1 = -1.10$ (True effect of STR on Test Score)
- $\beta_2 = -0.65$ (True effect of % EL on Test Score)
- $\delta_1 = 1.81$ (Effect of % EL on STR)

True Model:

$$\widehat{\text{Test Score}} = \underset{(7.41)}{686.03} - \underset{(0.38)}{1.10}\,STR - \underset{(0.04)}{0.65}\,\%\ EL$$

Omitted Regression:

$$\widehat{\text{Test Score}} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\,STR$$

Auxiliary Regression:

$$\widehat{\%\ EL} = \underset{(9.16)}{-19.85} + \underset{(0.46)}{1.81}\,STR$$

- Omitted Regression estimate for α_1 on STR is -2.28.

$$\alpha_1 = \beta_1 + \underbrace{\beta_2 \delta_1}_{bias}$$

- $\beta_1 = -1.10$ (True effect of STR on Test Score)
- $\beta_2 = -0.65$ (True effect of % EL on Test Score)
- $\delta_1 = 1.81$ (Effect of % EL on STR)
- So, for the **ommited regression**:

$$\begin{aligned}\alpha_1 &= -1.10 + (-0.65)(1.81)\\ &= -2.28\end{aligned}$$

True Model:

$$\widehat{\text{Test Score}} = 686.03 - 1.10\,STR - 0.65\,\%\ EL$$
$$(7.41) \quad (0.38) \quad (0.04)$$

Omitted Regression:

$$\widehat{\text{Test Score}} = 698.93 - 2.28\,STR$$
$$(9.47) \quad (0.48)$$

Auxiliary Regression:

$$\widehat{\%\ EL} = -19.85 + 1.81\,STR$$
$$(9.16) \quad (0.46)$$

- Omitted Regression estimate for α_1 on STR is -2.28.

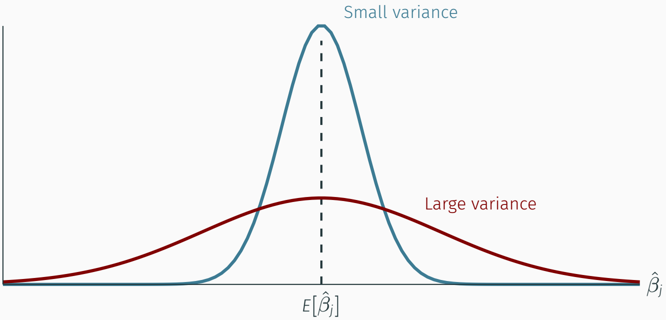
$$\alpha_1 = \beta_1 + \underbrace{\beta_2\delta_1}_{bias}$$

- $\beta_1 = -1.10$ (True effect of STR on Test Score)
- $\beta_2 = -0.65$ (True effect of % EL on Test Score)
- $\delta_1 = 1.81$ (Effect of % EL on STR)
- So, for the **ommited regression**:

$$\begin{aligned}\alpha_1 &= -1.10 + (-0.65)(1.81)\\ &= -2.28\end{aligned}$$

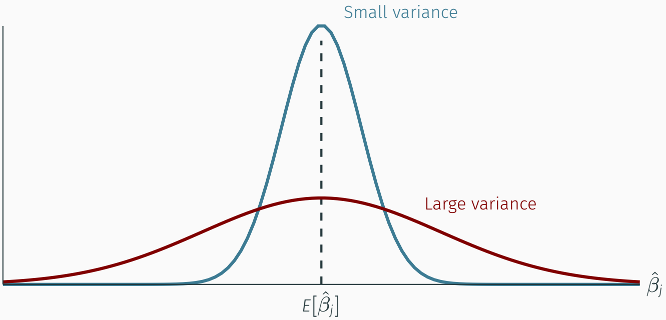
- The **bias** is $(-0.65)(1.81) = -1.18$

- How precise is our estimate $\hat{\beta}_j$?



Precision of OLS Estimates

- How precise is our estimate $\hat{\beta}_j$?
- We can talk of the variance, $var(\hat{\beta}_j)$ or the standard error, $SE(\hat{\beta}_j)$ of $\hat{\beta}_j$



- The variance of $\hat{\beta}_j$ is

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

compare with what we learned in Lecture 8

- The variance of $\hat{\beta}_j$ is

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

compare with what we learned in Lecture 8

- The standard error of $\hat{\beta}_j$ is simply the square root of the variance

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

Variance of $\hat{\beta}_1$

- The variance of $\hat{\beta}_j$ is

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

compare with what we learned in Lecture 8

- The standard error of $\hat{\beta}_j$ is simply the square root of the variance

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

- The **new term** in front is called the Variance Inflation Factor (VIF), explained below

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

- Variance is now affected by four things

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

- Variance is now affected by four things
    1. **Goodness of fit of the model:** SER

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

- Variance is now affected by four things
    1. **Goodness of fit of the model:** SER
        - Larger SER, larger $var(\hat{\beta}_j)$

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

- Variance is now affected by four things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_j)$
  2. **Sample size,** n

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

- Variance is now affected by four things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_j)$
  2. **Sample size,** n
     - Larger n, lower $var(\hat{\beta}_j)$

What Affects Variance

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

- Variance is now affected by four things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_j)$
  2. **Sample size,** n
     - Larger n, lower $var(\hat{\beta}_j)$
  3. **Variation in** X

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

- Variance is now affected by four things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_j)$
  2. **Sample size,** n
     - Larger n, lower $var(\hat{\beta}_j)$
  3. **Variation in** X
     - Larger $var(X)$, smaller $var(\hat{\beta}_j)$

What Affects Variance

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

- Variance is now affected by four things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_j)$
  2. **Sample size,** n
     - Larger n, lower $var(\hat{\beta}_j)$
  3. **Variation in** X
     - Larger $var(X)$, smaller $var(\hat{\beta}_j)$
  4. **Variance Inflation Factor (VIF),** $\frac{1}{(1-R_j^2)}$

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

- Variance is now affected by four things
  1. **Goodness of fit of the model:** SER
     - Larger SER, larger $var(\hat{\beta}_j)$
  2. **Sample size,** n
     - Larger n, lower $var(\hat{\beta}_j)$
  3. **Variation in** X
     - Larger $var(X)$, smaller $var(\hat{\beta}_j)$
  4. **Variance Inflation Factor (VIF),** $\frac{1}{(1-R_j^2)}$
     - Larger VIF, larger $var(\hat{\beta}_j)$

HOOD COLLEGE

- Two *independent* variables are multicollinear

$$corr(X_j, X_l) \neq 0 \text{ for } j \neq l$$

- Two *independent* variables are multicollinear

$$corr(X_j, X_l) \neq 0 \text{ for } j \neq l$$

- Multicollinearity between X variables does *not bias* OLS estimates

- Two *independent* variables are multicollinear

$$corr(X_j, X_l) \neq 0 \text{ for } j \neq l$$

- Multicollinearity between X variables does *not bias* OLS estimates
    - Remember, we pulled another variable out of ϵ into the regression

- Two *independent* variables are multicollinear

$$corr(X_j, X_l) \neq 0 \text{ for } j \neq l$$

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

- Two *independent* variables are multicollinear

$$corr(X_j, X_l) \neq 0 \text{ for } j \neq l$$

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

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

VIF and Multicollinearity II

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)

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

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

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)

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

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

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

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)

Example

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

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

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

$$R_1^2 \text{ for } X_{1i} = \gamma + \gamma X_{2i} + \gamma X_{3i}$$

HOOD COLLEGE

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)

Example

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

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

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

$$R_1^2 \text{ for } X_{1i} = \gamma + \gamma X_{2i} + \gamma X_{3i}$$
$$R_2^2 \text{ for } X_{2i} = \zeta_0 + \zeta_1 X_{1i} + \zeta_2 X_{3i}$$

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)

Example

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

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

- 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 III

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

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)

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

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)
- The R_j^2's tell us how much *other* regressors explain regressor X_j

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

- R_j^2 is the R^2 from an auxiliary regression of X_j on all other regressors (X's)
- The R_j^2's tell 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

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

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

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

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

- VIF quantifies the factor by which $var(\hat{\beta}_j)$ increases because of multicollinearity

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

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

- VIF quantifies the factor by which $var(\hat{\beta}_j)$ increases because of multicollinearity
- Baseline: $R_j^2 = 0 \implies$ *no* multicollinearity $\implies$ VIF = 1 (no inflation)

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

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

- VIF quantifies the factor by which $var(\hat{\beta}_j)$ increases because of multicollinearity
- Baseline: $R_j^2 = 0 \implies$ *no* multicollinearity $\implies$ VIF = 1 (no inflation)
- Larger $R_j^2 \implies$ larger VIF

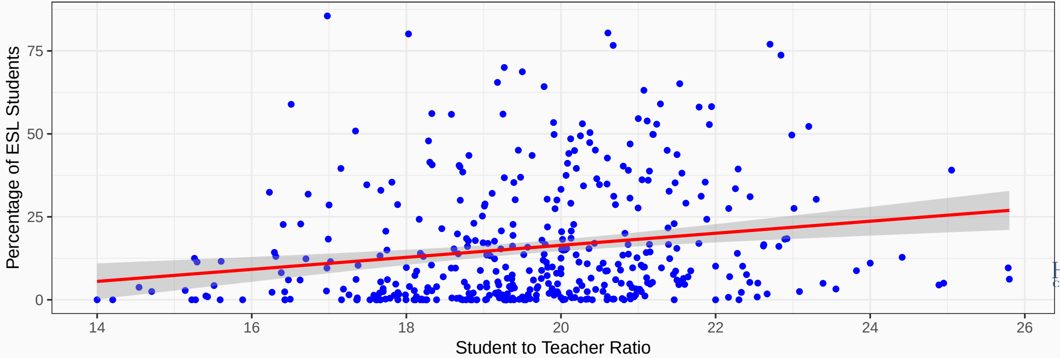
- Common to calculate the Variance Inflation Factor (VIF) for each regressor:

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

- VIF quantifies the factor by which $var(\hat{\beta}_j)$ increases because of multicollinearity
- 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 worrisome

```
xs.scatter<-ggplot(data=CASchool, aes(x=str,y=el_pct))+
  geom_point(color="blue")+
  geom_smooth(method="lm", color="red")+
  xlab("Student to Teacher Ratio")+ylab("Percentage of ESL Students")

xs.scatter
```



```r
library("car") # package for VIF function

# syntax: vif(lm.object)

vif(multireg) # "multireg" is our multivariate regression from before
```

```
##      str   el_pct
## 1.036495 1.036495
```

VIF and Multicollinearity: In R

```r
library("car") # package for VIF function

# syntax: vif(lm.object)

vif(multireg) # "multireg" is our multivariate regression from before
```

```
##      str   el_pct
## 1.036495 1.036495
```

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

Let's calculate it manually

```
auxreg<-lm(str~el_pct, data=CASchool)
summary(auxreg)
```

```
##
## Call:
## lm(formula = str ~ el_pct, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3343 -1.1313  0.0299  1.1296  6.3453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.33432    0.11993 161.212  < 2e-16 ***
## el_pct       0.01941    0.00497   3.906  0.00011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.86 on 418 degrees of freedom
## Multiple R-squared:  0.03521,    Adjusted R-squared:  0.0329
## F-statistic: 15.25 on 1 and 418 DF,  p-value: 0.0001095
```

```r
# r saves R^2, among many other things in the lm regression object saved
aux.r2<-summary(auxreg)$r.squared # save the auxiliary R^2 as aux.r2
our.vif<-1/(1-aux.r2) # VIF formula
our.vif
```

```
## [1] 1.036495
```

Example
For our Test Scores and Class Size example, what about district expenditures per student?

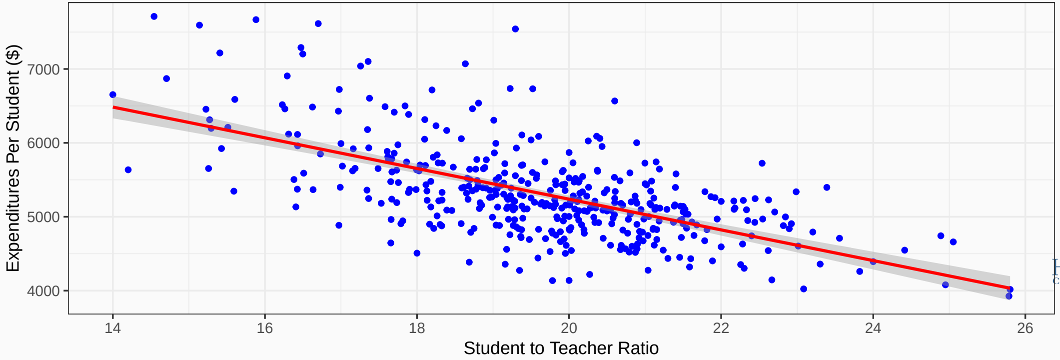
```
# reselect data to include expn too
CAcorr2<-subset(CASchool, select=c("testscr", "str", "expn_stu"))

# Make a correlation table
corr2<-cor(CAcorr2)
library("stargazer")
stargazer(corr2, type="latex", header=FALSE, float=FALSE)
```

	testscr	str	expn_stu
testscr	1	-0.226	0.191
str	-0.226	1	-0.620
expn_stu	0.191	-0.620	1

```
exp.scatter<-ggplot(data=CASchool, aes(x=str,y=expn_stu))+
  geom_point(color="blue")+
  geom_smooth(method="lm", color="red")+
  xlab("Student to Teacher Ratio")+ylab("Expenditures Per Student ($)")

exp.scatter
```



Example

1. $corr(\text{Test score}, \text{expn}) \neq 0$

Example

1. $corr(\text{Test score,expn}) \neq 0$
2. $corr(\text{STR,expn}) \neq 0$

Example

1. $corr(\text{Test score,expn}) \neq 0$
2. $corr(\text{STR,expn}) \neq 0$

- Omitting *expn* will **bias** $\hat{\beta}_1$ on STR

Example

1. $corr(\text{Test score,expn}) \neq 0$
2. $corr(\text{STR,expn}) \neq 0$

- Omitting *expn* will **bias** $\hat{\beta}_1$ on STR
- *Including expn* will **not** bias $\hat{\beta}_1$ on STR, but *will* make it less precise (higher variance)

Example

1. $corr(\text{Test score,expn}) \neq 0$
2. $corr(\text{STR,expn}) \neq 0$

- Omitting *expn* will **bias** $\hat{\beta}_1$ on STR
- *Including expn* will **not** bias $\hat{\beta}_1$ on STR, but *will* make it less precise (higher variance)
  - Data tells us little about the effect of a change in *STR* holding *expn* constant

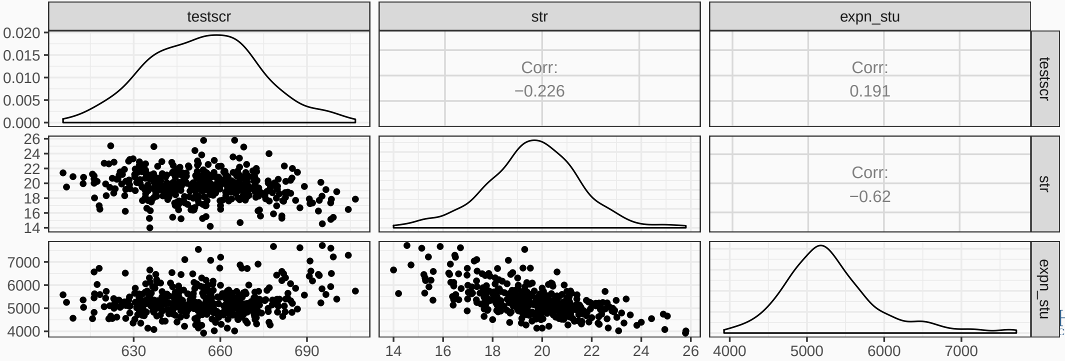
Example

1. $corr(\text{Test score,expn}) \neq 0$
2. $corr(\text{STR,expn}) \neq 0$

- Omitting *expn* will **bias** $\hat{\beta}_1$ on STR
- *Including expn* will **not** bias $\hat{\beta}_1$ on STR, but *will* make it less precise (higher variance)
  - Data tells us little about the effect of a change in *STR* holding *expn* constant
  - Hard to know what happens to test scores when high *STR* AND high *expn* and vice versa (*they rarely happen simultaneously*)!

Some Great Diagnostic Tools in R

```r
library("GGally") # see https://ggobi.github.io/ggally/
ggpairs(CAcorr2)
```



Multicollinearity Increases Variance

```r
expreg<-lm(testscr~str+expn_stu, data=CASchool)
summary(expreg)
```

```
##
## Call:
## lm(formula = testscr ~ str + expn_stu, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.507 -14.403   0.407  13.195  48.392
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 675.577174  19.562222  34.535   <2e-16 ***
## str          -1.763216   0.610914  -2.886   0.0041 **
## expn_stu      0.002487   0.001823   1.364   0.1733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.56 on 417 degrees of freedom
## Multiple R-squared:  0.05545,    Adjusted R-squared:  0.05092
## F-statistic: 12.24 on 2 and 417 DF,  p-value: 6.824e-06
```

```r
vif(expreg)
```

```
##      str expn_stu
## 1.624373 1.624373
```

- Including **`expn_stu`** increases variance of $\hat{\beta}_1$ by 1.62 times

Multicollinearity Increases Variance

	Dependent variable:	
	Test Score	
	(1)	(2)
Student Teacher Ratio	-2.280^{***}	-1.763^{***}
	(0.480)	(0.611)
Expenditures/Student		0.002
		(0.002)
Constant	698.933^{***}	675.577^{***}
	(9.467)	(19.562)
Observations	420	420
R^2	0.051	0.055
Adjusted R^2	0.049	0.051
Residual Std. Error	18.581 (df = 418)	18.562 (df = 417)
F Statistic	22.575^{***} (df = 1; 418)	12.241^{***} (df = 2; 417)
Note:		*p<0.1; **p<0.05; ***p<0.01

- We can see $SE(\hat{\beta}_1)$ on `str` increases from 0.480 to 0.611 when we add `expn_stu`

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

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

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

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

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

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

Perfect Multicollinearity

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

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

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

- $corr\big(\text{temperature (F), temperature (C)}\big) = 1$
- $R^2_j = 1$ is implying $VIF = \frac{1}{1-1}$ and $var\big(\hat{\beta}_j\big) = 0!$
- This is fatal for a regression

Perfect Multicollinearity

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

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

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

- $corr\big(\text{temperature (F), temperature (C)}\big) = 1$
- $R_j^2 = 1$ is implying $VIF = \frac{1}{1-1}$ and $var\big(\hat{\beta}_j\big) = 0!$
- This is fatal for a regression
    - A logical impossiblity, almost always caused by human error

Example

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

- %EL: the percentage of students learning English

Example

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

- %EL: the percentage of students learning English
- %ES: the percentage of students fluent in English

Example

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

- %EL: the percentage of students learning English
- %ES: the percentage of students fluent in English
- $ES = 100 - EL$

Example

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

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

Perfect Multicollinearity Example II

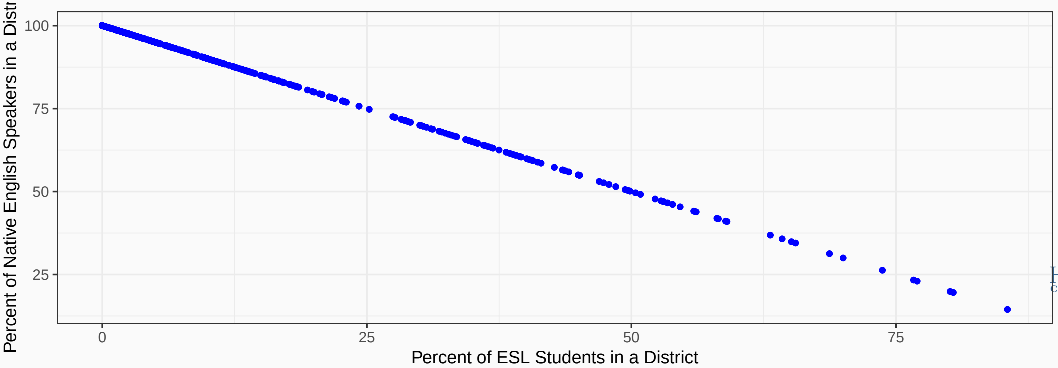
```r
# generate %EF variable from %EL
CASchool$ef_pct<-100-CASchool$el_pct

cor(CASchool$ef_pct, CASchool$el_pct)
```

```
## [1] -1
```

Perfect Multicollinearity Example III

```
mcol.scatter<-ggplot(CASchool, aes(x=el_pct,y=ef_pct))+
  geom_point(color="blue")+
  xlab("Percent of ESL Students in a District")+ylab("Percent of Native English
mcol.scatter
```



Perfect Multicollinearity Example III

```r
# try to run regression with both %EF and %EL
mcreg<-lm(testscr~str+el_pct+ef_pct, data=CASchool)
summary(mcreg)
```

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

- Note R ignores one of the perfectly multicollinear regressors (`ef_pct`)

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
    1. $corr(Y, Z) \neq 0$

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
    1. $corr(Y, Z) \neq 0$
    2. $corr(X_j, Z) \neq 0$

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
    1. $corr(Y, Z) \neq 0$
    2. $corr(X_j, Z) \neq 0$
    - If Z is *included* and X_j is collinear with Z, this does *not* cause a bias

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
  1. $corr(Y, Z) \neq 0$
  2. $corr(X_j, Z) \neq 0$
  - 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 of estimate:

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

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
  1. $corr(Y, Z) \neq 0$
  2. $corr(X_j, Z) \neq 0$
     - 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 of estimate:

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

- VIF from multicollinearity: $\frac{1}{(1-R_j^2)}$

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
  1. $corr(Y, Z) \neq 0$
  2. $corr(X_j, Z) \neq 0$
     - 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 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

HOOD COLLEGE

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
  1. $corr(Y, Z) \neq 0$
  2. $corr(X_j, Z) \neq 0$
  - 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 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

Bias and Precision of $\hat{\beta}_j$ (Summary)

- $\hat{\beta}_j$ on X_j is biased only if there is a variable (Z) omitted that:
  1. $corr(Y, Z) \neq 0$
  2. $corr(X_j, Z) \neq 0$
  - 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 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

- Again, how well does a linear model fit the data?

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

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

$$Y_i = \hat{Y}_i + \hat{\epsilon}_i$$

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

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

$$SER = \frac{SSE}{n - k - 1}$$

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

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

$$SER = \frac{SSE}{n - k - 1}$$

- A measure of the spread of the observations around the regression line (in units of Y), the average "size" of the residual

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

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

$$SER = \frac{SSE}{n - k - 1}$$

- 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[4]

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

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSE}{TSS} = (r_{X,Y})^2$$

- Again, R^2 is the fraction of the variation of the model ($\hat{Y}_i$) ("explained") to the variation of observations of Y_i ("total")

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

- Problem: R^2 of a regression increases *every* time a new variable is added (reduces SSE)

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

- Problem: R^2 of a regression increases *every* time a new variable is added (reduces SSE)
- This does *not* mean adding a variable improves the fit of the model per se, R^2 gets **inflated**

- Problem: R^2 of a regression increases *every* time a new variable is added (reduces SSE)
- 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 R^2:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Problem: R^2 of a regression increases *every* time a new variable is added (reduces SSE)
- 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 R^2:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- There are different methods to compute $\bar{R}^2$, and in the end, recall R^2 was never very useful, so don't worry about knowing the formula

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Note that $\frac{n-1}{n-k-1}$ is always greater than 1, so $\bar{R}^2 < R^2$

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Note that $\frac{n-1}{n-k-1}$ is always greater than 1, so $\bar{R}^2 < R^2$
- Adding a variable has two effects on $\bar{R}^2$:

HOOD COLLEGE

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Note that $\frac{n-1}{n-k-1}$ is always greater than 1, so $\bar{R}^2 < R^2$
- Adding a variable has two effects on $\bar{R}^2$:

  1. SSE falls, increasing $\bar{R}^2$

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Note that $\frac{n-1}{n-k-1}$ is always greater than 1, so $\bar{R}^2 < R^2$
- Adding a variable has two effects on $\bar{R}^2$:

  1. SSE falls, increasing $\bar{R}^2$
  2. $\frac{n-1}{n-k-1}$ increases (by increasing k)

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Note that $\frac{n-1}{n-k-1}$ is always greater than 1, so $\bar{R}^2 < R^2$
- Adding a variable has two effects on $\bar{R}^2$:
    1. SSE falls, increasing $\bar{R}^2$
    2. $\frac{n-1}{n-k-1}$ increases (by increasing k)
- $\bar{R}^2$ will change depending on which effect is larger

HOOD COLLEGE

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Note that $\frac{n-1}{n-k-1}$ is always greater than 1, so $\bar{R}^2 < R^2$
- Adding a variable has two effects on $\bar{R}^2$:

  1. SSE falls, increasing $\bar{R}^2$
  2. $\frac{n-1}{n-k-1}$ increases (by increasing k)

- $\bar{R}^2$ will change depending on which effect is larger
  - $\bar{R}^2$ could be negative!

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSE}{TSS}$$

- Note that $\frac{n-1}{n-k-1}$ is always greater than 1, so $\bar{R}^2 < R^2$
- Adding a variable has two effects on $\bar{R}^2$:
  1. SSE falls, increasing $\bar{R}^2$
  2. $\frac{n-1}{n-k-1}$ increases (by increasing k)
- $\bar{R}^2$ will change depending on which effect is larger
  - $\bar{R}^2$ could be negative!
  - Large sample sizes (n) make R^2 and $\bar{R}^2$ very close

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

```r
summary(reg)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

- Base R^2 (R calls it `multiple R-squared`) went up

```r
summary(multireg)
```

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

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

```r
summary(reg)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

- Base R^2 (R calls it `multiple R-squared`) went up
- `Adjusted R-squared` went down

```r
summary(multireg)
```

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