# PLSC 503 – Spring 2020
# Variable Selection and Specification Bias

March 3, 2020

Requires that:

- $\text{Cov}(\mathbf{X}, \mathbf{u}) = 0$, and
- the distribution of $\mathbf{X}$ does not depend on either $\boldsymbol{\beta}$ or $\sigma^2$.

"Truth":

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

Fitted model:

$$Y_i = \gamma_0 + \gamma_1 X_{1i} + e_i$$

Then:

$$e_i = \beta_2 X_{2i} + u_i$$

$$
\begin{aligned}
\mathsf{E}(e) &= \mathsf{E}(\beta_2 X_2 + u) \\
&= X_2 \mathsf{E}(\beta_2) + \mathsf{E}(u) \\
&\neq 0
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{E}(\gamma_1) &= \beta_1 + \frac{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2}\beta_2 \\
&= \beta_1 + b_{X_2 X_1}\beta_2
\end{aligned}
$$

where $b_{X_2 X_1}$ is the "slope" coefficient one obtains from regressing $X_2$ on $X_1$.

# Omitted Variable Bias, continued

If $\text{Cov}(X_1, X_2) = 0$ then

- $E(\hat{\gamma}_1) = \beta_1$, but
- $E(\hat{\gamma}_0) \neq \beta_0$.

If $\text{Cov}(X_1, X_2) \neq 0$ then

- $E(\hat{\gamma}_1) \neq \beta_1$ and $E(\hat{\gamma}_0) \neq \beta_0$
- In the simple bivariate case,
    - if $Cov(X_1, X_2) > 0$ then $E(|\hat{\gamma}_1|) > |\beta_1|$,
    - if $Cov(X_1, X_2) < 0$ then $E(|\hat{\gamma}_1|) < |\beta_1|$.

Recall that for one $X$:

$$\widehat{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

and for two $X$s:

$$\widehat{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2(1 - R^2_{X_1 X_2})}$$

Also, because $\hat{e}_i \neq \hat{u}_i$,

$$\text{E}(\sigma_e^2) = \sigma_u^2 + f(\beta_2, X_1) \leftarrow \text{Bias}$$

For the “true” DGP

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

and fitted model

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{e}$$

where $\mathbf{Z} \subset \mathbf{X}$, we have

$$\begin{aligned} \boldsymbol{\Gamma} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \end{aligned}$$

and so

$$\begin{aligned} \mathrm{E}(\boldsymbol{\Gamma}) &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}\boldsymbol{\beta}. \end{aligned}$$

Now assume a "true" model:

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

and fitted model:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{e}$$

where $\mathbf{X} \subset \mathbf{Z}$. This means:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\Theta} + \mathbf{u}$$

where $\boldsymbol{\Theta} = 0$.

Results:

- $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $E(\hat{\sigma^2}) = \sigma^2$, but
- $\widehat{\text{Var}(\boldsymbol{\beta})} > \text{Var}(\boldsymbol{\beta}) \;\leftarrow\;$ Inefficiency

Implication: *Pre-Test Bias*

# Omitted Variable Bias: Simulated Example

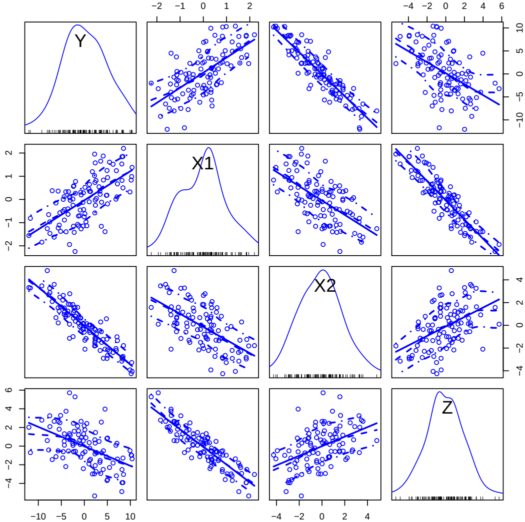
"True" model:

$$Y_i = 0 + 1.0X_{1i} - 2.0X_{2i} + u_i$$

Simulation:

```
> N <- 100
> X1<-rnorm(N)                # <- X1
> X2<-(-X1)+1.5*(rnorm(N)) # <- correlated w/X1
> Y<-X1-(2*X2)+(2*(rnorm(N))) # <- Y
Z<- (-2*X1) + rnorm(N)  # <- correlated w/X1 but irrelevant
> data <- data.frame(Y=Y,X1=X1,X2=X2,Z=Z)
```

# Scatterplot Matrix



# Correctly Specified Model

```
> correct<-lm(Y~X1+X2)
> summary(correct)

Residuals:
   Min     1Q Median     3Q    Max
-5.721 -1.209  0.093  1.198  5.915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03311    0.21249  -0.156  0.87651
X1           0.81690    0.26718   3.057  0.00288 **
X2          -2.13652    0.13844 -15.433  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.116 on 97 degrees of freedom
Multiple R-squared:  0.8295,Adjusted R-squared:  0.826
F-statistic:   236 on 2 and 97 DF,  p-value: < 2.2e-16
```

# Overspecified Model

```
> overspec<-lm(Y~X1+X2+Z)
> summary(overspec)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9809 -1.0442 -0.0265  1.2609  6.0201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01570    0.21420  -0.073  0.94173
X1           0.82148    0.26785   3.067  0.00281 **
X2          -2.11735    0.14105 -15.011  < 2e-16 ***
Z            0.01662    0.02202   0.755  0.45220
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.12 on 96 degrees of freedom
Multiple R-squared:  0.8306,Adjusted R-squared:  0.8253
F-statistic: 156.8 on 3 and 96 DF,  p-value: < 2.2e-16
```

# Underspecified Model

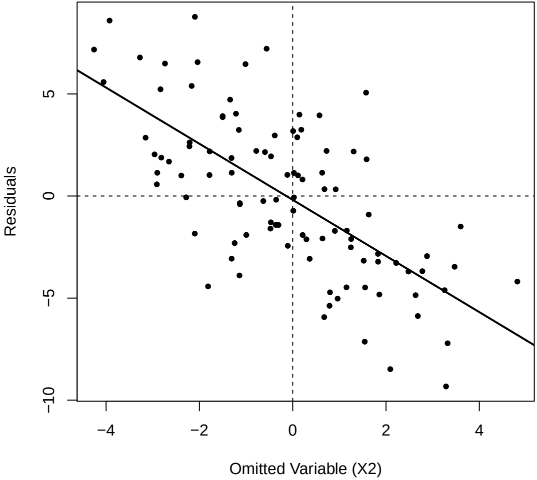
```
> incorrect<-lm(Y~X1)
> summary(incorrect)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3297 -2.9762 -0.0672  2.4828  8.7787

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2704     0.3913   0.691    0.491
X1            3.2783     0.3964   8.270 6.71e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.913 on 98 degrees of freedom
Multiple R-squared:  0.411,Adjusted R-squared:  0.405
F-statistic: 68.39 on 1 and 98 DF,  p-value: 6.714e-13
```

# Omitted Variable Plot



**Nothing Beats a Good Theory. Period.**

Also:

- "Model specification tests" $\leftarrow$ meh
- Examine residuals
- Proxy variables...
- *Resist the urge to overspecify!*