BIOS7345 Lab 4

Sarah Lotspeich

28 September 2018

Underfitting

Generate data

Simulate `n = 1000` independent observations from the following distributions:

1. $x_{1i} \sim N(0, 1)$

```
x1 <- rnorm(n = 1000, mean = 0, sd = 1)
```

2. $X_{2i} \sim N(x_{1i}, 1)$

```
x2 <- rnorm(n = 1000, mean = x1, sd = 1)
```

3. $y_i \sim N(2 + 2x_{1i} + 3x_{2i}, 1)$

```
y <- rnorm(n = 1000, mean = 2 + 2*x1 + 3*x2, sd = 1)
```

Based on the way we've generate our data, the correct model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i.$$

Begin by fitting the **correct** model,

```
model_full <- ols(y ~ x1 + x2)
```

and then fit the reduced (e.g. underfitted) model for y_i on x_{1i} without adjusting for x_{2i}.

```
model_red <- ols(y ~ x1)
```

Why is this model underfitted?

Because from the way we generated the y_i, we know that y_i depends on not only x_{1i} but also x_{2i}.

Compare the coefficients on x_{1i} from the full ($\hat{\beta}_1$)

```
model_full$coefficients["x1"]
```

```
##       x1
## 2.040604
```

and reduced ($\hat{\beta}_1^*$) models.

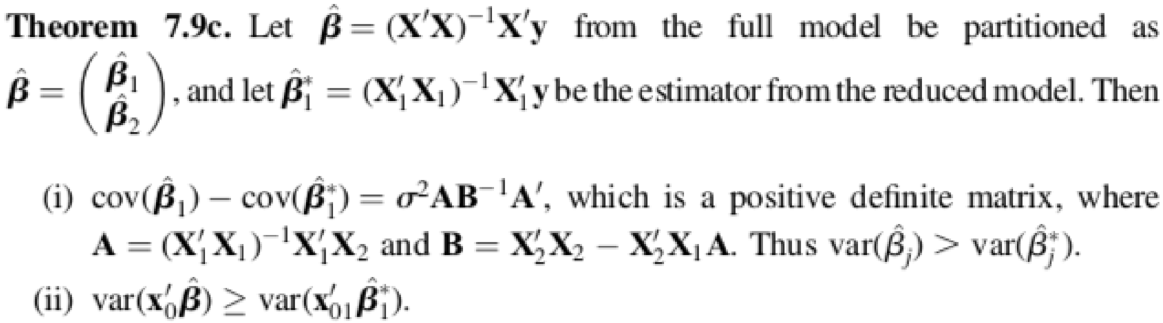
```
model_red$coefficients["x1"]
```

```
##       x1
## 5.255582
```

Is $\hat{\beta}_1^*$ biased?

Yes, we see that $\hat{\beta}_1^*$ is inflated over $\hat{\beta}_1$ (from the true model).

Compare the standard errors of the coefficients on x_{1i} from the full ($SE(\hat{\beta}_1)$)



Theorem 7.9c. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ from the full model be partitioned as $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$, and let $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ be the estimator from the reduced model. Then

(i) $\text{cov}(\hat{\boldsymbol{\beta}}_1) - \text{cov}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2\mathbf{A}\mathbf{B}^{-1}\mathbf{A}'$, which is a positive definite matrix, where $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ and $\mathbf{B} = \mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1\mathbf{A}$. Thus $\text{var}(\hat{\beta}_j) > \text{var}(\hat{\beta}_j^*)$.

(ii) $\text{var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) \geq \text{var}(\mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*)$.

Figure 1: From Rencher

```r
diag(model_full$var)["x1"] %>% sqrt()
```

```
##         x1
## 0.04787001
```

and reduced ($SE(\hat{\beta}_1^*)$) models.

```r
diag(model_red$var)["x1"] %>% sqrt()
```

```
##        x1
## 0.1001313
```

Is $SE(\hat{\beta}_1^*)$ biased?

> Yes, once again we see that $SE(\hat{\beta}_1^*)$ is inflated over $SE(\hat{\beta}_1)$. Thus, we have that the reduced model has positively biased estimators and standard errors.

Using the full and reduced model, show that Theorem 7.9c (i) holds:

```r
#LHS: cov(beta-hat1) - cov(beta-hat1*)
lhs <- diag(model_full$var)["x1"] - diag(model_red$var)["x1"]

#RHS: sigma2A(B-inv)A'
A <- solve(t(x1)%*%x1)%*%t(x1)%*%x2
B <- t(x2)%*%x1-t(x2)%*%x1%*%A
sigma2 <- var(y)
rhs <- sigma2*A%*%solve(B)%*%t(A)

#Check LHS = RHS
lhs == rhs
```

```
##       [,1]
## [1,] FALSE
```

What do we observe?

> Theorem 7.9c holds true iff we know the true error variance, σ^2. In this case we don't! So we can only plug in estimates, which can be biased. Thus, the theorem doesn't hold with estimates (only true values) and we cannot necessarily conclude that the true $Var(\hat{\beta}_1) > Var(\hat{\beta}_1^*)$.

Compare the estimated variances.

```r
#estimated variance from full model (correct)
model_full$stats["Sigma"]
```

```
##    Sigma
## 1.025279
```

```r
#estimated variance from reduced model (underfitted)
model_red$stats["Sigma"]
```

```
##    Sigma
## 3.148531
```

Is $\hat{s}^{2*}$ biased?

> We see that the estimated variance of the reduced model was greater than that of the full (true) model. This is **Theorem 7.9d.**.

Fit the model

$$x_{2i} = \gamma_0 + \gamma_1 x_{1i} + \delta_i,$$

where $\delta_i \sim N(0, \theta)$.

```r
model_btwn <- ols(x2 ~ x1)
```

We know that $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ (correct model) and $x_{2i} = \gamma_0 + \gamma_1 x_{1i} + \delta_i$ (between model). If we begin with the correct model, we have

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2[\gamma_0 + \gamma_1 x_{1i} + \delta_i] + \epsilon_i \text{ subbing in the between model} \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)x_{1i} + (\beta_2\delta_i + \epsilon_i) \end{aligned}$$

Now if we set this to be equal to the reduced model,

$$\begin{aligned} y_i &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)x_{1i} + (\beta_2\delta_i + \epsilon_i) \\ \beta_0^* + \beta_1^* x_{1i} + \epsilon_i^* &\overset{set}{=} (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)x_{1i} + (\beta_2\delta_i + \epsilon_i) \end{aligned}$$

Hence, $\beta_1^* = \beta_1 + \beta_2\gamma_1$. Use this relationship to get β_1 based only on your reduced and between models

```r
model_red$coefficients["x1"] - model_full$coefficients["x2"]*model_btwn$coefficients["x1"]
```

```
##       x1
## 2.040604
```

and compare it to $\hat{\beta}_1$ from the full model.

```r
model_full$coefficients["x1"]
```

```
##       x1
## 2.040604
```

Mediator variables

Suppose x_1 represents a measure of how a person was parented as a child, and researchers want to know whether this affects how confident a person feels about parenting their own children (y_i).

It is believed that the way in which a person is parented affects their self confidence and self-esteem later in life (x_2), which in turn affects how confident a person feels about parenting their own children (y_i), i.e. x_2 is a mediator of the relationship between x_1 and y.

There are also other indirect effects of x_1 on y through other unmeasured mechanisms (e.g. parenting strategies).

Suppose one fits the above reduced model. Is $\hat{\beta}_1^*$ still biased?

> It's not biased because we are interested in the effect of x_1 on y. If we put x_2 into the analysis, since x_2 is a mediator, it will take away most effect on y from x_1.

Overfitting

Generate data

Simulate `n = 1000` independent observations from the following distributions:

1. $x_{1i} \sim N(0, 1)$

```
x1 <- rnorm(n = 1000, mean = 0, sd = 1)
```

2. $X_{2i} \sim N(x_{1i}, 1)$

```
x2 <- rnorm(n = 1000, mean = x1, sd = 1)
```

3. $y_i \sim N(2 + 2x_{1i} + 3x_{2i}, 1)$

```
y <- rnorm(n = 1000, mean = 2 + 2*x1, sd = 1)
```

Based on the way we've generate our data, the correct model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i.$$

Begin by fitting the full (e.g. overfitted) model,

```
model_full <- ols(y ~ x1 + x2)
```

and then fit the reduced (**correct**) model for y_i on x_{1i} without adjusting for x_{2i}.

```
model_red <- ols(y ~ x1)
```

Why is this model overfitted?

> Because from the way we generated the y_i, we know that y_i depends only on x_{1i} but we controlled for x_{2i} anyway.

Compare the coefficients on x_{1i} from the full ($\hat{\beta}_1$)

```
model_full$coefficients["x1"]
```

```
##       x1
## 2.034473
```

and reduced ($\hat{\beta}_1^*$) models.

```
model_red$coefficients["x1"]
```

```
##       x1
## 2.064601
```

Is $\hat{\beta}_1^*$ biased?

> No, $\hat{\beta}_1^* \approx \hat{\beta}_1$ because β_2 is almost 0 (since y_i doesn't depend on x_2 by design).

Compare the standard errors of the coefficients on x_{1i} from the full ($SE(\hat{\beta}_1)$)

```
diag(model_full$var)["x1"] %>% sqrt()
```

```
##         x1
## 0.04467672
```

and reduced ($SE(\hat{\beta}_1^*)$) models.

```
diag(model_red$var)["x1"] %>% sqrt()
```

```
##         x1
## 0.03169523
```

Is $SE(\hat{\beta}_1^*)$ biased?

> Yes, we see that $SE(\hat{\beta}_1^*)$ from the correct model is smaller than $SE(\hat{\beta}_1)$ from the overfitted model. Thus, we have that overfitted model has unbiased estimates but inflated standard errors.

Compare the estimated variances.

```
#estimated variance from full model (correct)
model_full$stats["Sigma"]
```

```
##    Sigma
## 1.000674
```

```
#estimated variance from reduced model (underfitted)
model_red$stats["Sigma"]
```

```
##    Sigma
## 1.000631
```

Is $\hat{s}^{2*}$ biased?

> No, the estimated variance from the overfitted model is not biased.

Independent covariates

Repeat the overfit and underfit exercises (above) by simulating $x_{2i} \sim N(0, 1)$ instead (e.g. x_{2i} is not correlated with x_{1i}) and see how the results differ.

Underfitting

```
#generate new data
x1 <- rnorm(n = 1000, mean = 0, sd = 1)
x2 <- rnorm(n = 1000, mean = x1, sd = 1)
y <- rnorm(n = 1000, mean = 2 + 2*x1 + 3*x2, sd = 1)

#fit full model y ~ x1 + x2
model_full <- ols(y ~ x1 + x2)
#fit reduced model y ~ x1
model_red <- ols(y ~ x1)

#compare coefficients on x1
model_full$coefficients["x1"]
```

```
##        x1
## 2.046419
```

```
model_red$coefficients["x1"]
```

```
##        x1
## 5.132051
```

```
#compare SEs on coefficients on x1
diag(model_full$var)["x1"] %>% sqrt()
```

```
##         x1
## 0.04516813
```

```
diag(model_red$var)["x1"] %>% sqrt()
```

```
##        x1
## 0.1025886
```

```
#compare estimated variances
model_full$stats["Sigma"]
```

```
##    Sigma
## 1.016645
```

```
model_red$stats["Sigma"]
```

```
##    Sigma
## 3.268353
```

Observations on overfitting when $x_1 \perp x_2$:

1. Coefficient from reduced model > coefficient from full (correct) model
2. Standard error from reduced model > standard error from full (correct) model
3. Variance greater for reduced than full model

Overfitting

```
#generate new data
x1 <- rnorm(n = 1000, mean = 0, sd = 1)
x2 <- rnorm(n = 1000, mean = 0, sd = 1)
y <- rnorm(n = 1000, mean = 2 + 2*x1, sd = 1)

#fit full model y ~ x1 + x2
model_full <- ols(y ~ x1 + x2)
#fit reduced model y ~ x1
model_red <- ols(y ~ x1)

#compare coefficients on x1
model_full$coefficients["x1"]
```

```
##        x1
## 2.061815
```

```
model_red$coefficients["x1"]
```

```
##        x1
## 2.058939
```

```r
#compare SEs on coefficients on x1
diag(model_full$var)["x1"] %>% sqrt()
```

```
##         x1
## 0.03192189
```

```r
diag(model_red$var)["x1"] %>% sqrt()
```

```
##         x1
## 0.03189016
```

```r
#compare estimated variances
model_full$stats["Sigma"]
```

```
##     Sigma
## 0.9959488
```

```r
model_red$stats["Sigma"]
```

```
##     Sigma
## 0.9966466
```

Observations on overfitting when $x_1 \perp x_2$:

1. Coefficient from reduced model = coefficient from full (correct) model
2. Standard error from reduced model = standard error from full (correct) model
3. Same Variance for reduced and full models