Stat Theory Homework 2.

Bohao Tang

1: (i) Correctly specified means that under true distribution P there exist $\beta_0, \beta_1 \in R$ such that

$$E_P(Y|X) = \beta_0 + \beta_1 X$$

Misspecified means there are no such β_0, β_1

For example if $X \in R$ and $Y = 1_{\{X<0\}}$

then (X, Y) can not be described by this model since $E(Y|X) = 1_{\{X<0\}} = 0, \text{ or } 1$ but $\beta_0 + \beta_1 X$ can only have one value β_0 or the whole real line for its range.

Consider $V = \text{span}\{1, X\}$ then the model constrains parts that orthogonal to V and not constrain parts in V.

(ii)

$$\hat\beta = \left[\begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}' \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}\right]^{-1} \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}' \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \overline{X} \\ \overline{X} & \overline{X^2} \end{pmatrix}^{-1} \begin{pmatrix} \overline{Y} \\ \overline{XY} \end{pmatrix} = \frac{1}{\overline{X^2} - (\overline{X})^2} \begin{pmatrix} \overline{X^2} & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix} \begin{pmatrix} \overline{Y} \\ \overline{XY} \end{pmatrix}$$

$$= \frac{1}{\overline{X^2} - (\overline{X})^2} \begin{pmatrix} \overline{X^2}\,\overline{Y} - \overline{X}\,\overline{XY} \\ \overline{XY} - \overline{X}\,\overline{Y} \end{pmatrix}$$

By WLLN, we know that

$$\hat{\beta} \xrightarrow{P} \frac{1}{EX^2 - (EX)^2} \begin{pmatrix} EX^2 EY - EXEXY \\ EXY - EXEY \end{pmatrix}$$

$$= \frac{1}{varX} \begin{pmatrix} varX\, EY - cov(X,Y) EX \\ cov(X,Y) \end{pmatrix} = \beta^*$$

For limit distribution of $\sqrt{n}(\hat{\beta} - \beta^*)$, we use delta method, we know by CLT that

$$\sqrt{n} \left\{ \begin{pmatrix} \bar{X} \\ \bar{Y} \\ \overline{XY} \\ \overline{X^2} \end{pmatrix} - \begin{pmatrix} EX \\ EY \\ EXY \\ EX^2 \end{pmatrix} \right\} \xrightarrow{d} N\left(0, \underbrace{\begin{bmatrix} varX & cov(X,Y) & cov(X,XY) & cov(X,X^2) \\ \vdots & varY & cov(Y,XY) & cov(Y,X^2) \\ \vdots & \vdots & varXY & cov(XY, X^2) \\ \vdots & \vdots & \vdots & varX^2 \end{bmatrix}}_{= D}\right)$$

therefore $\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, V)$

where $V = wDw^T$

where $w = \begin{pmatrix} \frac{\partial f}{\partial x_1}, & \frac{\partial f}{\partial x_2}, & \frac{\partial f}{\partial x_3}, & \frac{\partial f}{\partial x_4} \\ \frac{\partial g}{\partial x_1}, & \frac{\partial g}{\partial x_2}, & \frac{\partial g}{\partial x_3}, & \frac{\partial g}{\partial x_4} \end{pmatrix}$ $x_1 = EX$, $x_2 = EY$, $x_3 = EXY$, $x_4 = EX^2$

where $f = \frac{x_4 x_2 - x_1 x_3}{x_4 - x_1^2}$, $g = \frac{x_3 - x_1 x_2}{x_4 - x_1^2}$

(iii) if $X \perp\!\!\!\perp Y$, $cov(X,Y) = 0 \Rightarrow \beta_1^* = 0$

(iv) see the code part.

(v) $\hat{\beta}_{OLS} \xrightarrow{P} \beta^*$ in the setting $(X_i, Y_n) \overset{i.i.d}{\sim} P$

We compare it with estimator $\breve{\beta}$ where

$$\breve{\beta}\left[(X_i, Y_i)_{i=1}^{n}\right] = \begin{pmatrix} \bar{Y} - \hat{\beta}_1 \bar{X} \\ \breve{\beta}_1 \end{pmatrix}$$

where $\breve{\beta}_1 = \text{Median}\left\{ \dfrac{Y_{2k} - Y_{2k-1}}{X_{2k} - X_{2k-1}} \right\}_{k=1}^{\lfloor \frac{n}{2} \rfloor}$

Then see the coding part.

2: We need to prove that if there exists two sequences of random vectors U_n, V_n such that

$\exists\, C < +\infty \quad P(\|U_n\| > C) \to 0$

$V_n / \|Y_n\| \xrightarrow{P} \vec{0}$

and $Y_n = U_n + V_n$

Then $\exists\, C' < +\infty \quad P(\|Y_n\| > C') \to 0$

Proof: Let $C' = 2C$ then $P(\|Y_n\| > C') = P(2\|Y_n\| > C' + \|Y_n\|)$

$= P(\|Y_n\| > \frac{C'}{2} + \frac{1}{2}\|Y_n\|) = P[\|U_n + V_n\| > C + \frac{1}{2}\|Y_n\|]$

$\leq P(\|U_n\| > C) + P(\|V_n\| > \frac{1}{2}\|Y_n\|) \to 0$

therefore $Y_n = O_p(1)$

Coding Part

Bohao Tang

April 1, 2018

iv

Two distribution is

$$Y = 1 + 2X + \epsilon$$

and

$$Y = 1 + 2X^2 + \epsilon$$

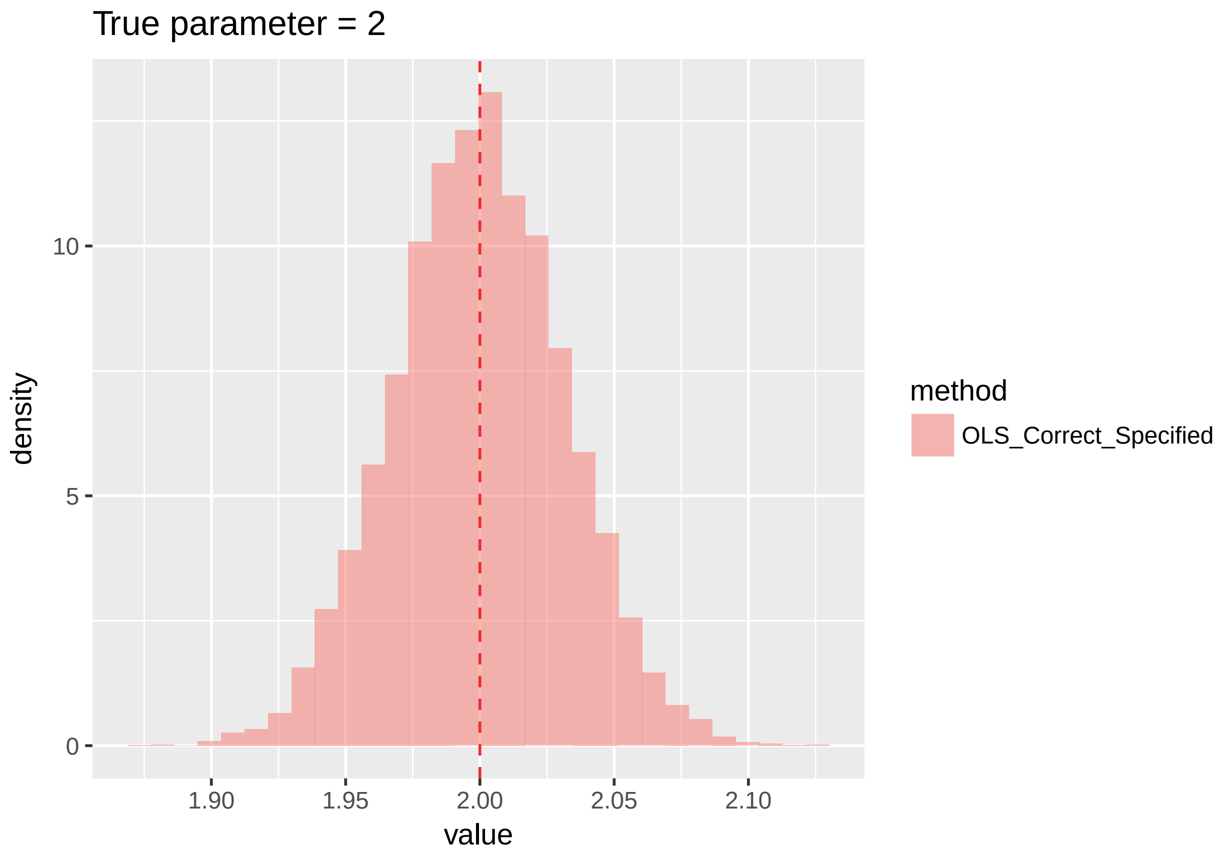
Where X, ϵ mutual independent and $\epsilon, X \sim N(0, 1)$.

Then the regression $E(Y|X) = \beta_0 + \beta_1 X$ correctly specifies first distribution and misspecifies the second.

```
model1 = function(X, eps){
    1 + 2*X + eps
}
model2 = function(X, eps){
    1 + 2 * X^2 + eps
}
```

For the first situation, we run estimation 10000 times, each have 1000 samples and only focus on β_1. The red dashed line is the true β_1:

```
library(ggplot2)

beta1 = c()

n = 1000
for(i in 1:10000){
    X = rnorm(n, 0, 1)
    eps = rnorm(n, 0, 1)

    Y = model1(X, eps)

    estimator = cov(X, Y) / var(X)

    beta1 = c(beta1, estimator)
}

data = data.frame(value = beta1)
data$method = "OLS_Correct_Specified"

ggplot(data, aes(value, fill = method)) +
    geom_histogram(alpha = 0.5, aes(y = ..density..), position = 'identity') +
    geom_vline(xintercept = 2, linetype="dashed", color="firebrick2") +
    ggtitle("True parameter = 2")
```



Then the bias and variance for $\hat{\beta}_1$ is:

```
bias = mean(beta1) - 2
bias
```

```
## [1] -9.523031e-05
```

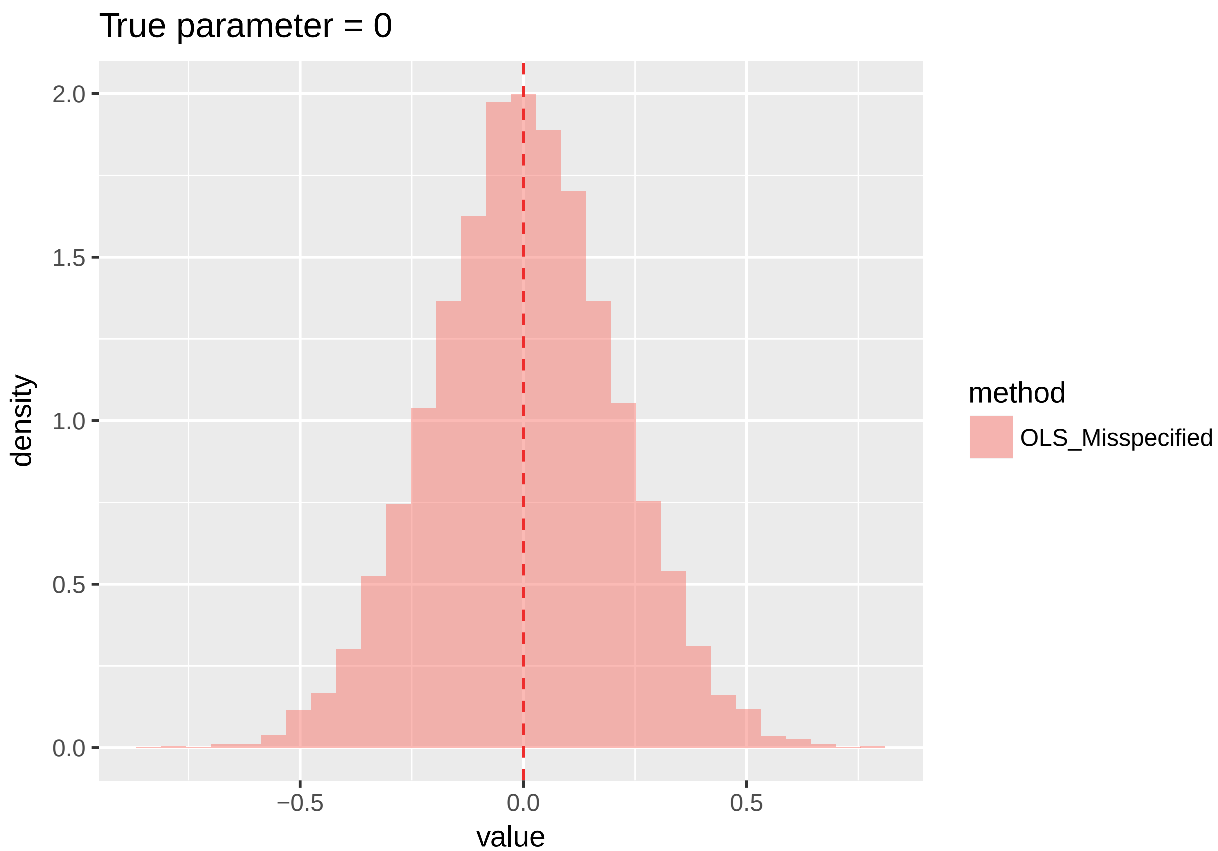
```
variance = var(beta1)
variance
```

```
## [1] 0.0009997933
```

In the second situation, now true β_1 doesn't exit, we need to compare the result to $\beta_1^* = \frac{Cov(X,Y)}{Var(X)} = 0$. We do the same thing as above:

```
beta1mis = c()

n = 1000
for(i in 1:10000){
    X = rnorm(n, 0, 1)
    eps = rnorm(n, 0, 1)

    Y = model2(X, eps)

    estimator = cov(X, Y) / var(X)

    beta1mis = c(beta1mis, estimator)
}
```

```r
datamis = data.frame(value = beta1mis)
datamis$method = "OLS_Misspecified"

ggplot(datamis, aes(value, fill = method)) +
    geom_histogram(alpha = 0.5, aes(y = ..density..), position = 'identity') +
    geom_vline(xintercept = 0, linetype="dashed", color="firebrick2") +
    ggtitle("True parameter = 0")
```



Then the bias and variance for $\hat{\beta}_1^*$ is:

```r
bias = mean(beta1mis) - 0
bias
```

```
## [1] 0.00111693
```

```r
variance = var(beta1mis)
variance
```

```
## [1] 0.04135902
```

v

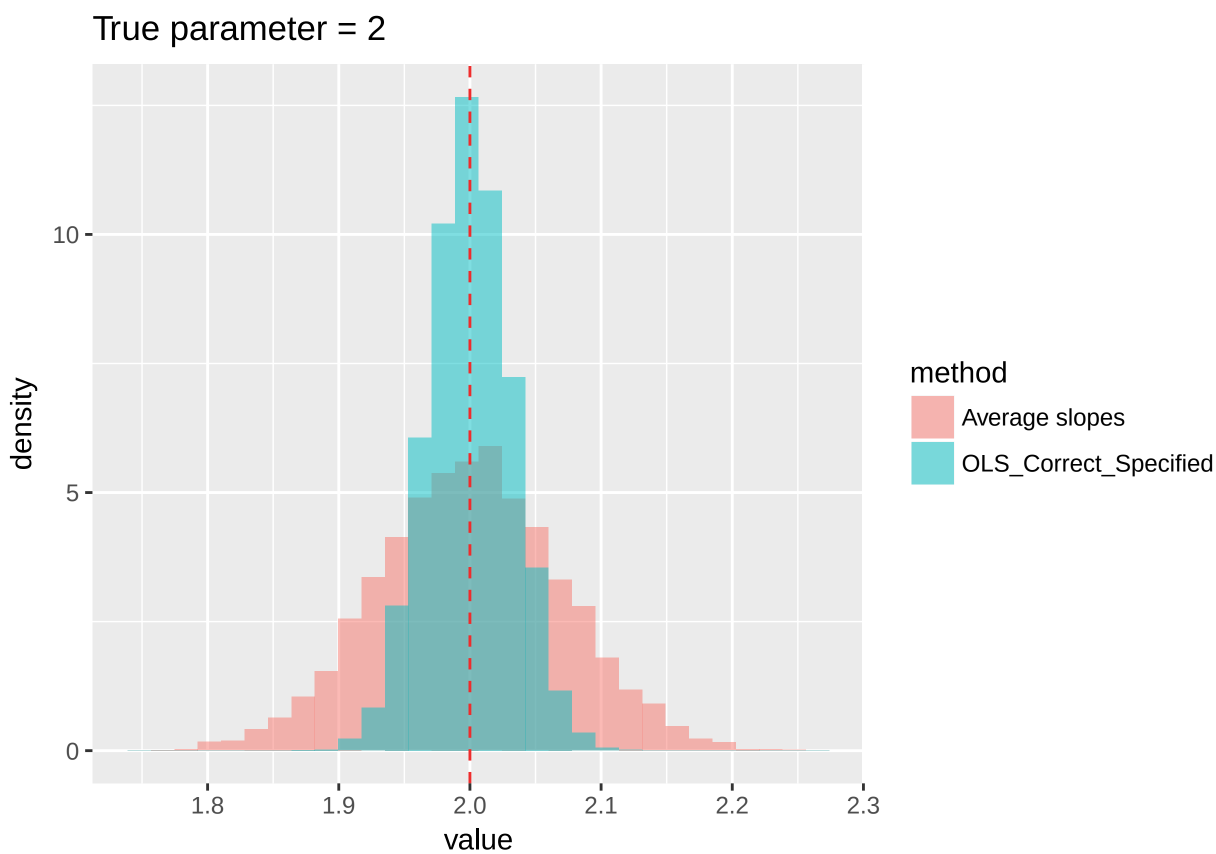
We do the model:

$$Y = 1 + 2X + \epsilon$$

where X, ϵ mutual independent and $\epsilon, X \sim N(0, 1)$.

And the second estimator mentioned in homework solution, which is just a median of slopes. We use data from `iv` for OLS estimator and for second estimator:

```
slope = c()

n = 1000
index = 1:1000
oddi = index[index%%2 == 1]
eveni = index[index%%2 == 0]
for(i in 1:10000){
    X = rnorm(n, 0, 1)
    eps = rnorm(n, 0, 1)

    Y = model1(X, eps)

    estimator =
        median( (Y[eveni] - Y[oddi]) / (X[eveni] - X[oddi]) )

    slope = c(slope, estimator)
}

slopes = data.frame(value = slope)
slopes$method = "Average slopes"

estimators = rbind(data, slopes)

ggplot(estimators, aes(value, fill = method)) +
    geom_histogram(alpha = 0.5, aes(y = ..density..), position = 'identity') +
    geom_vline(xintercept = 2, linetype="dashed", color="firebrick2") +
    ggtitle("True parameter = 2")
```



Therefore we can see OLS is a better estimator for β_1.