Hw3

Hongkai Lou, Shubham Kumar, Moyi Li, Brinda Narayanan

2024-10-16

```r
library(car)
```

```
## Loading required package: carData
```

```r
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```r
auto <- read.table('auto.txt', header = T)
lm1 <- lm(log(mpg) ~ cylinders+log(displacement) + log(weight) + year, data = auto)
summary(lm1)
```

```
##
## Call:
## lm(formula = log(mpg) ~ cylinders + log(displacement) + log(weight) +
##     year, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42025 -0.06517  0.00495  0.06552  0.41614
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.272094   0.370167  19.645   <2e-16 ***
## cylinders         -0.011463   0.010370  -1.105    0.270
## log(displacement) -0.049557   0.046485  -1.066    0.287
## log(weight)       -0.788979   0.062716 -12.580   <2e-16 ***
## year               0.031889   0.001694  18.826   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1163 on 392 degrees of freedom
## Multiple R-squared:  0.8842, Adjusted R-squared:  0.883
## F-statistic: 748.3 on 4 and 392 DF,  p-value: < 2.2e-16
```

1(a): The fraction of variance explained by the predictors is 0.8842, meaning 88.42% of the variance in log(mpg) is explained by the predictors. The remaining 11.58% of the variance is unexplained by this model. log(weight) and year are the most significant variables.

1(b):

```
vif(lm1)
```

```
##          cylinders log(displacement)       log(weight)              year
##           9.113381         17.895468          9.090709          1.143479
```

We have very severe muitlicollinearity in this model, especially around variables log(displacement), and substantial multicollinearity log(weight), and cylinder. These variables might be closely related and we need to address this issue.

1(c):

```
lm2 <- lm(log(mpg) ~ cylinders+log(displacement)  + year, data = auto)
summary(lm2)
```

```
##
## Call:
## lm(formula = log(mpg) ~ cylinders + log(displacement) + year,
##     data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53993 -0.07077  0.00711  0.07554  0.56358
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.203717   0.213118  15.033   <2e-16 ***
## cylinders         -0.006191   0.012261  -0.505    0.614
## log(displacement) -0.462198   0.038976 -11.859   <2e-16 ***
## year               0.030258   0.001998  15.141   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1376 on 393 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8362
## F-statistic: 674.9 on 3 and 393 DF,  p-value: < 2.2e-16
```

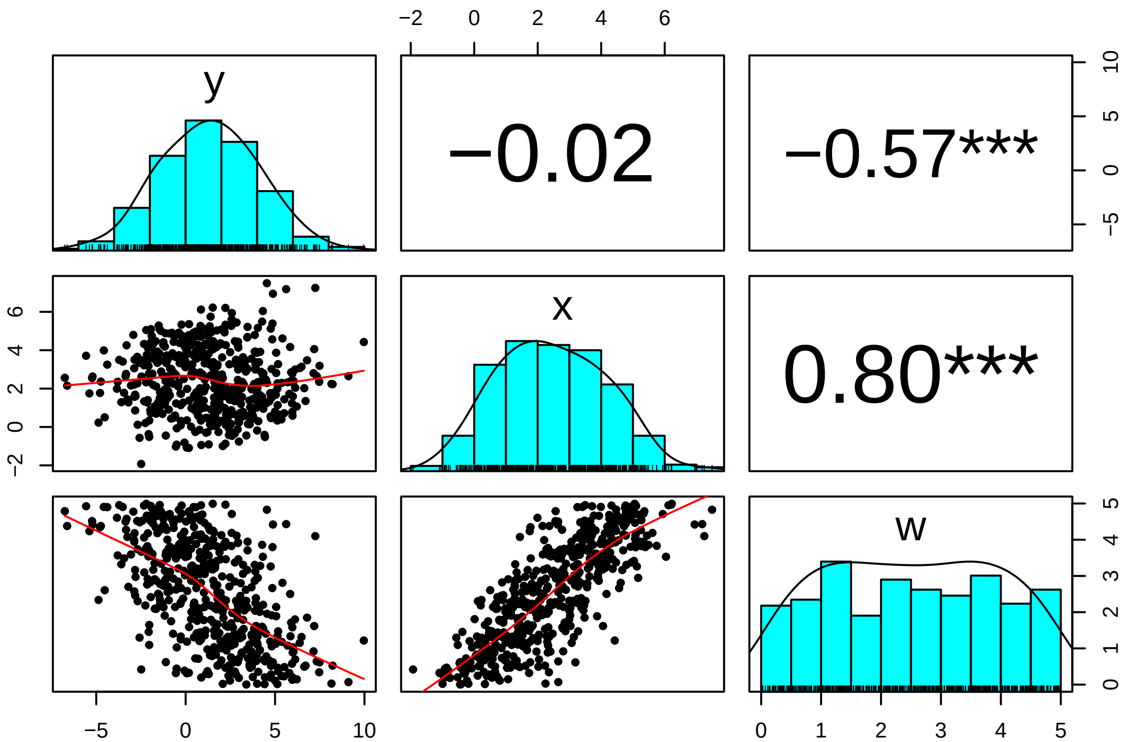
We see that the coefficient of cylinders increases by 0.005, which is about 50 percent of the original estimate. However, log(displacement)'s estimate coefficient decreases about 9.4 times. The difference in the estimated coefficient compared to previous model is so dramatic, indicating a very strong correlation between log(displacement and) from the theorem Omitted variable bias.

```
n = 500
w <- runif(n, 0, 5)
epislon <- rnorm(n,0,1)
error <- rnorm(n,0,1)
x = w+epislon
y = 4+2*x-3*w+error
```

2(a): This case is fork. In fork case, we need to include control to block back-door path. So if w -> y and w ->x->y, we need to include w in our case to study x->y.

2(b) and 2(c)

```
dataset2 <- data.frame(y, x, w)
pairs.panels(dataset2, ellipses=F, stars=T)
```




```r
summary(dataset2)
```

```
##        y                  x                 w
##  Min.   :-6.7570   Min.   :-1.926   Min.   :0.002471
##  1st Qu.:-0.6792   1st Qu.: 1.193   1st Qu.:1.317741
##  Median : 1.3413   Median : 2.348   Median :2.521267
##  Mean   : 1.3126   Mean   : 2.447   Mean   :2.517728
##  3rd Qu.: 3.1717   3rd Qu.: 3.698   3rd Qu.:3.724157
##  Max.   : 9.9639   Max.   : 7.493   Max.   :4.992563
```

```r
for(i in 1:3){
  print(sd(dataset2[, i]))
}
```

```
## [1] 2.749537
## [1] 1.693881
## [1] 1.417844
```

```r
lm3 <- lm(y~x, data = dataset2)
summary(lm3)
```

```
##
## Call:
## lm(formula = y ~ x, data = dataset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0652 -1.9795  0.0412  1.8410  8.7245
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.40304    0.21638   6.484 2.15e-10 ***
```

```
## x            -0.03697    0.07272  -0.508    0.611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.752 on 498 degrees of freedom
## Multiple R-squared:  0.0005188,  Adjusted R-squared:  -0.001488
## F-statistic: 0.2585 on 1 and 498 DF,  p-value: 0.6114
```

The coefficient of x is 0.08163, with a t-statistics of 1.286 and p-value of 0.199. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since $0.199 > 0.05$, we fail to reject the null hypothesis, meaning that the coefficient of x is not significant at the 0.05 level. The 95% confidence interval is

```r
c(0.08163 - qt(0.975, 497)*0.06347, 0.08163 + qt(0.975, 497)*0.06347 )
```

```
## [1] -0.04307259  0.20633259
```

Which did not cover the true slope for x.

2(d):

```r
lm4 <- lm(y ~ x+w, data = dataset2)
summary(lm4)
```

```
##
## Call:
## lm(formula = y ~ x + w, data = dataset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5206 -0.6654 -0.0116  0.7122  3.0056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.04395    0.09220   43.86   <2e-16 ***
## x            1.99360    0.04477   44.53   <2e-16 ***
## w           -3.02285    0.05349  -56.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.011 on 497 degrees of freedom
## Multiple R-squared:  0.8654, Adjusted R-squared:  0.8649
## F-statistic:  1598 on 2 and 497 DF,  p-value: < 2.2e-16
```

The coefficient of x is 2.07140, with a t-statistics of 46.29 and p-value less than 2e-16. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since $2e\text{-}16 < 0.05$, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The 95% confidence interval is

```r
c(2.07140 - qt(0.975, 497)*0.04475, 2.07140 + qt(0.975, 497)*0.04475 )
```

```
## [1] 1.983477 2.159323
```

Which cover the true slope for x, which is 2.

2(e): The VIF score is moderate in this case

```r
vif(lm4)
```

```
##        x        w
## 2.809319 2.809319
```

3(a): This is a collider case

3(b):

```
set.seed(005536893)
x <- runif(n,0,5)
y <- x+rnorm(n,0,1)
w <- 2*x+3*y+4+rnorm(n,0,1)
dataset3 <- data.frame(x,y,w)
round(cor(dataset3), 3)
```

```
##       x     y     w
## x 1.000 0.847 0.932
## y 0.847 1.000 0.971
## w 0.932 0.971 1.000
```

```
summary(dataset3)
```

```
##        x                  y                 w
##  Min.   :0.00599   Min.   :-2.689   Min.   :-3.409
##  1st Qu.:1.14523   1st Qu.: 1.106   1st Qu.: 9.983
##  Median :2.57510   Median : 2.652   Median :16.731
##  Mean   :2.49278   Mean   : 2.548   Mean   :16.642
##  3rd Qu.:3.85535   3rd Qu.: 3.907   3rd Qu.:23.269
##  Max.   :4.99558   Max.   : 7.057   Max.   :34.616
```

```
for(i in 1:3){
  print(sd(dataset3[, i]))
}
```

```
## [1] 1.483955
## [1] 1.788945
## [1] 8.090725
```

3(c):

```
lmcollider1 <- lm(y~x, data = dataset3)
summary(lmcollider1)
```

```
##
## Call:
## lm(formula = y ~ x, data = dataset3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4091 -0.6449 -0.0166  0.6748  2.7855
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.001168   0.083200   0.014    0.989
## x           1.021506   0.028687  35.609   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9509 on 498 degrees of freedom
## Multiple R-squared:  0.718,  Adjusted R-squared:  0.7174
## F-statistic:  1268 on 1 and 498 DF,  p-value: < 2.2e-16
```

The coefficient of x is 1.021506, with a t-statistics of 35.609 and p-value less than 2e-16. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since 2e-16 < 0.05, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The 95% confidence interval is

```
c(1.021506 - qt(0.975, 497)*0.028687, 1.021506 + qt(0.975, 497)*0.028687 )
```

```
## [1] 0.9651433 1.0778687
```

Which covers the true coefficient of $B_1 = 1$

3(d):

```
lmcollider2 <- lm(y~x+w, data = dataset3)
summary(lmcollider2)
```

```
##
## Call:
## lm(formula = y ~ x + w, data = dataset3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03572 -0.20999  0.01996  0.21922  0.84185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.213212   0.033551  -36.16   <2e-16 ***
## x           -0.529567   0.026199  -20.21   <2e-16 ***
## w            0.305302   0.004805   63.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3152 on 497 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.969
## F-statistic:  7790 on 2 and 497 DF,  p-value: < 2.2e-16
```

The coefficient of x is -0.529567, with a t-statistics of -20.31 and p-value less than 2e-16. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since 2e-16 < 0.05, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The 95% confidence interval is

```
c(-0.529567 - qt(0.975, 497)*0.026199, -0.529567 + qt(0.975, 497)*0.026199 )
```

```
## [1] -0.5810414 -0.4780926
```

Which unsurprisingly, does not actually covers the true coefficient of $B_1 = 1$. It captures the wrong one.

3(e): Although the VIF is still less than 10, it does imply substantial multicollinearity.

```
vif(lmcollider2)
```

```
##        x        w
## 7.593269 7.593269
```

3(f): The R squared value may indicate the second model is better than the first model, with an adjusted r-squared of 0.969 compared to 0.7174. However, this is not the right model, as the true relationship between x and y is not captured, where we end up getting a wrong estimate of coefficient of x with relation to y.

4(a): This is a pipe. case

4(b):

```
set.seed(005536893)
x <- runif(n,0,5)
```

```r
w <- x+rnorm(n,0,1)
y <- 2*w+rnorm(n,0,1)
dataset4 <- data.frame(x,y,w)
round(cor(dataset4), 3)
```

```
##       x     y     w
## x 1.000 0.829 0.847
## y 0.829 1.000 0.964
## w 0.847 0.964 1.000
```

```r
summary(dataset4)
```

```
##        x                  y                w
##  Min.   :0.00599   Min.   :-6.129   Min.   :-2.689
##  1st Qu.:1.14523   1st Qu.: 2.458   1st Qu.: 1.106
##  Median :2.57510   Median : 5.065   Median : 2.652
##  Mean   :2.49278   Mean   : 5.109   Mean   : 2.548
##  3rd Qu.:3.85535   3rd Qu.: 7.933   3rd Qu.: 3.907
##  Max.   :4.99558   Max.   :13.968   Max.   : 7.057
```

```r
for(i in 1:3){
  print(sd(dataset4[, i]))
}
```

```
## [1] 1.483955
## [1] 3.68711
## [1] 1.788945
```

4(c):

```r
lmpipe1 <- lm(y~x, data = dataset4)
summary(lmpipe1)
```

```
##
## Call:
## lm(formula = y ~ x, data = dataset4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5547 -1.5013  0.0378  1.4700  5.8438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02353    0.18075   -0.13    0.896
## x            2.05895    0.06232   33.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.066 on 498 degrees of freedom
## Multiple R-squared:  0.6867, Adjusted R-squared:  0.6861
## F-statistic:  1091 on 1 and 498 DF,  p-value: < 2.2e-16
```

The coefficient of x is 2.05895, with a t-statistics of 33.04 and p-value less than 2e-16. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since 2e-16 $<$ 0.05, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level.

4(d):

```r
lmpipe2 <- lm(y~x+w, data = dataset4)
summary(lmpipe2)
```

```
##
## Call:
## lm(formula = y ~ x + w, data = dataset4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.77755 -0.63372 -0.04721  0.62889  2.93638
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02577    0.08523  -0.302   0.7625
## x            0.10136    0.05534   1.832   0.0676 .
## w            1.91638    0.04590  41.749   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9741 on 497 degrees of freedom
## Multiple R-squared:  0.9305, Adjusted R-squared:  0.9302
## F-statistic:  3326 on 2 and 497 DF,  p-value: < 2.2e-16
```

The coefficient of x is 0.10136, with a t-statistics of 1.832 and p-value is equal to 0.0676. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since $0.0676 > 0.05$, we fail to reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level.

4(e): The adjusted R-squared value is higher for the second model, which is 0.9302 compared to 0.6861 from the first model. However, the first model is better since it demonstrate relationship between x and y, and in pipe case we should not control for w.

5(a)

```r
set.seed(1)
x1 = runif(500,0,4) # part a
x2 = 0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

The linear model has the form $y = B_1 * x_1 + B_2 * x_2 + \epsilon, with \epsilon ~ N(0, 1). B_1 = 2 and B_2 = 0.3$

5(b): The correlation between x1 and x2 is 0.8351212 5(c)

```r
cor(x1, x2)
```

```
## [1] 0.985686
```

```r
lm5 <- lm(y~x1)
summary(lm5)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.02857 -0.55855 -0.00726  0.71476  2.04095
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.07059    0.09416   21.99   <2e-16 ***
## x1           2.08373    0.04124   50.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 498 degrees of freedom
## Multiple R-squared:  0.8368, Adjusted R-squared:  0.8364
## F-statistic:  2553 on 1 and 498 DF,  p-value: < 2.2e-16
```

The coefficient of x_1 is 1.4396, with a t-statistics of 1.996 and p-value is equal to 0.0487. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since $0.0487 < 0.05$, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The coefficient of x_2 is 1.0097, with a t-statistics of 0.891 and p-value is equal to 0.3754. If we do a hypothesis test with $H_0 : B_2 = 0$ and $H_1 : B_2 \neq 0$, since $0.3754 > 0.05$, we fail reject the null hypothesis, meaning that the coefficient of x is not significant at the 0.05 level. The true coefficients are both covered for x_1 and x_2 in the CI at 95 confidence level, since the standard error is very big.

5(d):

```r
lm6 <- lm(y~x1)
summary(lm6)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.02857 -0.55855 -0.00726  0.71476  2.04095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.07059    0.09416   21.99   <2e-16 ***
## x1           2.08373    0.04124   50.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 498 degrees of freedom
## Multiple R-squared:  0.8368, Adjusted R-squared:  0.8364
## F-statistic:  2553 on 1 and 498 DF,  p-value: < 2.2e-16
```

The coefficient of x_1 in this model is 1.9759, with a t-statistics of 4.986 and p-value is equal to 2.66e-06. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since $2.66\text{e-}06 < 0.05$, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The estimate of coefficient is very close to 2, and with relatively large standard error, the true coefficient is included in the 95% CI.

5(e):

```r
lm7 <- lm(y~x2)
summary(lm7)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.1654 -0.7362 -0.1092  0.9038  2.6040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14077    0.09868   21.69   <2e-16 ***
## x2           4.05572    0.08545   47.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.099 on 498 degrees of freedom
## Multiple R-squared:  0.819,  Adjusted R-squared:  0.8186
## F-statistic:  2253 on 1 and 498 DF,  p-value: < 2.2e-16
```

The coefficient of x_2 in this model is 2.8996, with a t-statistics of 4.58 and p-value is equal to 1.37e-05. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since 1.37e-05 $<$ 0.05, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The estimate of coefficient is very close to 3, whereas our true coefficient is 0.3 for x2. If we calculate the confidence interval:

```r
confint(lm7)
```

```
##                2.5 %   97.5 %
## (Intercept) 1.946883 2.334652
## x2          3.887842 4.223604
```

We see that the true interval of x_2 is not covered in the confidence interval.

5(f): Not really. x_2 is a variable created by x_1 for relatively small values. When regressing y and both x1 and x2, variance can be explained from both variables. When we only regress y on x1, we would expect the estimate of coefficient to increase, account for the missing predictor variable x2. However, when we regress on x2, x2 will try to account for the unexplained variance from x1, since it is x1 -> x2. Thus, the coefficient of estimate will be far away from the true coefficient.

6

6. We want to show that if we fit $Y = B_0 + B_1 + \varepsilon$ when true model is $Y = B_0 + B_1 x_1 + B_2 x_2 + \varepsilon$.

$$E(\hat{B}_1) = B_1 + B_2 r \frac{S_2}{S_1}$$

Our model becomes $Y = B_0 + B_1 x_1 + (B_2 x_2 + \varepsilon)$

Then $\hat{B}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ $\quad y_i - \bar{y} = B_1(x_{i1} - \bar{x}_1) + B_2(x_{i2} - \bar{x}_2)$

$$= \frac{\sum (x_i - \bar{x})(B_1 x_1 + B_2 x_2 - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_{i1} - \bar{x}_1) \cdot B_1 (x_{i1} - \bar{x}_1) + \sum (x_{i1} - \bar{x}_1) \cdot B_2 (x_2 - \bar{x}_2)}{\sum (x_{i1} - \bar{x}_1)^2}$$

$$= B_1 + \frac{B_2 \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{S_{11}}$$

$$= B_1 + B_2 r \frac{S_2}{S_1}$$

The B_1 bias is 0 when r, the sample correlation coefficient between X_1 and X_2, is 0. Or $B_2 = 0$

$S_2 = 0$

7

7) $y = \beta_1 z_1 + \beta_2 z_2 + e$

$\left(\bar{z}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij} = 0,\ s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_{ij}-0)^2 = 1\right.$ and $r = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}z_{i2}$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

How does $VIF = 1/(1-r^2)$ affect Variance?

$$V(b) = \sigma^2 (Z^T Z)^{-1}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

$$Z^T Z = \begin{bmatrix} \sum z_{i1}^2 & \sum z_{i1}z_{i2} \\ \sum z_{i1}z_{i2} & \sum z_{i2}^2 \end{bmatrix}$$

$$\sum z_{i2}^2 = n-1,\quad \sum z_{i1}^2 = n-1$$

$$\sum z_{i1}z_{i2} = (n-1)r$$

$$Z^T Z = (n-1)\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

$$(Z^T Z)^{-1} = \frac{1}{(n-1)^2(1-r^2)}\begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}\cancel{(n-1)}$$

$$Var \rightarrow \sigma^2 \frac{1}{(n-1)(1-r^2)}\begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}$$

$$VIF = \frac{1}{1-r^2} \rightarrow \begin{array}{l} Var(\hat{\beta}_1) = \sigma^2 \frac{VIF}{n-1} \\ Var(\hat{\beta}_2) = \sigma^2 \frac{VIF}{n-1} \end{array}$$

VIF affects the variance by: b/c it is proportional to $\frac{1}{1-r^2}$, as the correlation between predictors increases, the variance of the slope increases

8: We would expect that as the amount of the crime goes up, the demand for the bikes will go down, because people will feel more dangerous when using or renting a bike. The type of crime would also affect the demand, as different types of crime will bring different kinds of danger. We would expect that for crimes happened more on streets and may bring more danger to citizens will have more negative impact, such as battery, assault, robbery, and homicide. On the other hand, criminals that involve home intrude or economic behavior, such as deceptive practice, criminal trespassing, theft, burglary will have less impact on demand of bikes. Further results will require investigations.

```
bike <- read.csv('bike.csv')
bike <- bike[, -c(1, 46)]
bike$assault_battery <- bike$ASSAULT+bike$BATTERY
lm_bike <- lm(trips ~ ASSAULT+ROBBERY+BURGLARY+THEFT+CRIMINAL_TRESPASS+NARCOTICS+HOMICIDE+BATTERY+DECEP
```

```r
summary(lm_bike)
```

```
##
## Call:
## lm(formula = trips ~ ASSAULT + ROBBERY + BURGLARY + THEFT + CRIMINAL_TRESPASS +
##     NARCOTICS + HOMICIDE + BATTERY + DECEPTIVE_PRACTICE, data = bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72133 -0.33453  0.03637  0.40328  1.20438
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.58787    0.44003  14.971  < 2e-16 ***
## ASSAULT             -1.05217    0.14602  -7.206 5.01e-12 ***
## ROBBERY             -0.29581    0.11073  -2.671 0.007979 **
## BURGLARY            -0.23686    0.06586  -3.597 0.000379 ***
## THEFT                0.92036    0.13586   6.774 6.96e-11 ***
## CRIMINAL_TRESPASS    0.10399    0.10006   1.039 0.299543
## NARCOTICS            0.04040    0.08268   0.489 0.625496
## HOMICIDE            -0.10997    0.06325  -1.739 0.083168 .
## BATTERY              0.33909    0.18709   1.812 0.070955 .
## DECEPTIVE_PRACTICE   0.32041    0.12790   2.505 0.012790 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5708 on 290 degrees of freedom
## Multiple R-squared:  0.6568, Adjusted R-squared:  0.6461
## F-statistic: 61.66 on 9 and 290 DF,  p-value: < 2.2e-16
```

```r
vif(lm_bike)
```

```
##            ASSAULT            ROBBERY           BURGLARY              THEFT
##          12.348550           6.797078           1.616178          12.887908
##  CRIMINAL_TRESPASS          NARCOTICS           HOMICIDE            BATTERY
##           7.766560           6.025335           2.141520          18.434109
## DECEPTIVE_PRACTICE
##          13.839823
```

The VIF of this model is relatively high. We remove the highly correlated variables in the next one.

```r
lm_bike_2 <- lm(trips ~ ASSAULT+ROBBERY+BURGLARY+THEFT+CRIMINAL_TRESPASS+NARCOTICS+HOMICIDE+STALKING, da
summary(lm_bike_2)
```

```
##
## Call:
## lm(formula = trips ~ ASSAULT + ROBBERY + BURGLARY + THEFT + CRIMINAL_TRESPASS +
##     NARCOTICS + HOMICIDE + STALKING, data = bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81903 -0.36008  0.01208  0.42149  1.44284
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.87599    0.38604  17.811  < 2e-16 ***
```

```
## ASSAULT           -0.90013    0.12474  -7.216 4.66e-12 ***
## ROBBERY           -0.25224    0.10694  -2.359  0.01899 *
## BURGLARY          -0.21969    0.06603  -3.327  0.00099 ***
## THEFT              1.17763    0.09168  12.846  < 2e-16 ***
## CRIMINAL_TRESPASS  0.18626    0.09502   1.960  0.05091 .
## NARCOTICS          0.07871    0.07619   1.033  0.30242
## HOMICIDE          -0.10988    0.06112  -1.798  0.07327 .
## STALKING           0.15227    0.06521   2.335  0.02022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5739 on 291 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6423
## F-statistic:  68.1 on 8 and 291 DF,  p-value: < 2.2e-16
```

```r
vif(lm_bike_2)
```

```
##           ASSAULT           ROBBERY          BURGLARY             THEFT
##          8.914489          6.270854          1.607178          5.805088
## CRIMINAL_TRESPASS         NARCOTICS          HOMICIDE          STALKING
##          6.927884          5.061134          1.978166          1.864677
```

We try to add other kinds of crime to observe the p-value in the next model.

```r
lm_bike_3 <- lm(trips ~ ASSAULT+ROBBERY+BURGLARY+THEFT+CRIMINAL_TRESPASS+NARCOTICS+HOMICIDE+STALKING+WEA
summary(lm_bike_3)
```

```
##
## Call:
## lm(formula = trips ~ ASSAULT + ROBBERY + BURGLARY + THEFT + CRIMINAL_TRESPASS +
##     NARCOTICS + HOMICIDE + STALKING + WEAPONS_VIOLATION + CRIM_SEXUAL_ASSAULT,
##     data = bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79127 -0.34625  0.03328  0.40527  1.33578
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.12118    0.43001  16.561  < 2e-16 ***
## ASSAULT             -0.85314    0.13134  -6.496   3.6e-10 ***
## ROBBERY             -0.25658    0.10659  -2.407 0.016700 *
## BURGLARY            -0.23178    0.06589  -3.518 0.000505 ***
## THEFT                1.10544    0.09723  11.369  < 2e-16 ***
## CRIMINAL_TRESPASS    0.15440    0.09551   1.616 0.107078
## NARCOTICS            0.09623    0.08185   1.176 0.240705
## HOMICIDE            -0.08277    0.06190  -1.337 0.182214
## STALKING             0.15945    0.06546   2.436 0.015464 *
## WEAPONS_VIOLATION   -0.10835    0.06690  -1.620 0.106420
## CRIM_SEXUAL_ASSAULT  0.14537    0.08501   1.710 0.088347 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5705 on 289 degrees of freedom
## Multiple R-squared:  0.6583, Adjusted R-squared:  0.6465
## F-statistic: 55.69 on 10 and 289 DF,  p-value: < 2.2e-16
```

```r
vif(lm_bike_3)
```

```
##             ASSAULT             ROBBERY            BURGLARY               THEFT
##           10.001706            6.304686            1.619377            6.608389
##   CRIMINAL_TRESPASS           NARCOTICS            HOMICIDE            STALKING
##            7.084412            5.912049            2.052986            1.901469
##   WEAPONS_VIOLATION CRIM_SEXUAL_ASSAULT
##            3.221232            3.650927
```

We want to observe the effect of none crime variables.

```r
lm_bike_4 <- lm(trips ~ ASSAULT+ROBBERY+BURGLARY+THEFT+CRIMINAL_TRESPASS+NARCOTICS+HOMICIDE+STALKING+WEA
summary(lm_bike_4)
```

```
##
## Call:
## lm(formula = trips ~ ASSAULT + ROBBERY + BURGLARY + THEFT + CRIMINAL_TRESPASS +
##     NARCOTICS + HOMICIDE + STALKING + WEAPONS_VIOLATION + EDU,
##     data = bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67189 -0.35608  0.03456  0.40394  1.38433
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.39338    0.40474  15.796  < 2e-16 ***
## ASSAULT           -0.74505    0.13256  -5.621 4.48e-08 ***
## ROBBERY           -0.22001    0.10555  -2.084 0.037996 *
## BURGLARY          -0.25057    0.06558  -3.821 0.000163 ***
## THEFT              1.07811    0.09464  11.392  < 2e-16 ***
## CRIMINAL_TRESPASS  0.19344    0.09461   2.045 0.041800 *
## NARCOTICS          0.08323    0.08056   1.033 0.302414
## HOMICIDE          -0.09130    0.06086  -1.500 0.134642
## STALKING           0.11769    0.06489   1.814 0.070747 .
## WEAPONS_VIOLATION -0.11110    0.06618  -1.679 0.094247 .
## EDU                0.86723    0.28262   3.069 0.002355 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5642 on 289 degrees of freedom
## Multiple R-squared:  0.6658, Adjusted R-squared:  0.6542
## F-statistic: 57.57 on 10 and 289 DF,  p-value: < 2.2e-16
```

```r
vif(lm_bike_4)
```

```
##           ASSAULT           ROBBERY          BURGLARY             THEFT
##         10.414198          6.319822          1.640280          6.400023
## CRIMINAL_TRESPASS         NARCOTICS          HOMICIDE          STALKING
##          7.105360          5.854622          2.028703          1.909894
## WEAPONS_VIOLATION               EDU
##          3.222103          1.208843
```

There are some other predictor variables that will increase r^2 values but the cost is it will highly impact the p-value and estimate of other important variables (crime variables) that we wish to study and conduct hypothesis test on. Thus, we will not include them in the model. An example of including Capacity in the linear regression is included below.

```
lm_bike_5 <- lm(trips ~ ASSAULT+ROBBERY+BURGLARY+THEFT+CRIMINAL_TRESPASS+NARCOTICS+HOMICIDE+STALKING+WE
summary(lm_bike_5)
```

```
##
## Call:
## lm(formula = trips ~ ASSAULT + ROBBERY + BURGLARY + THEFT + CRIMINAL_TRESPASS +
##     NARCOTICS + HOMICIDE + STALKING + WEAPONS_VIOLATION + EDU +
##     CAPACITY, data = bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70949 -0.29137  0.04395  0.36263  1.38321
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.060478   0.379951  15.951  < 2e-16 ***
## ASSAULT           -0.709088   0.123500  -5.742 2.38e-08 ***
## ROBBERY           -0.178654   0.098437  -1.815  0.07058 .
## BURGLARY          -0.148368   0.062896  -2.359  0.01900 *
## THEFT              0.961006   0.089785  10.703  < 2e-16 ***
## CRIMINAL_TRESPASS  0.089315   0.089404   0.999  0.31863
## NARCOTICS          0.102630   0.075045   1.368  0.17251
## HOMICIDE          -0.084349   0.056657  -1.489  0.13764
## STALKING           0.050866   0.061205   0.831  0.40662
## WEAPONS_VIOLATION -0.096784   0.061634  -1.570  0.11744
## EDU                0.700595   0.264221   2.652  0.00846 **
## CAPACITY           0.048575   0.007197   6.750 8.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5252 on 288 degrees of freedom
## Multiple R-squared:  0.7114, Adjusted R-squared:  0.7004
## F-statistic: 64.54 on 11 and 288 DF,  p-value: < 2.2e-16
```

```
vif(lm_bike_5)
```

```
##           ASSAULT           ROBBERY          BURGLARY             THEFT
##         10.433622          6.344404          1.741185          6.648246
## CRIMINAL_TRESPASS         NARCOTICS          HOMICIDE          STALKING
##          7.323390          5.863221          2.029374          1.961215
## WEAPONS_VIOLATION               EDU          CAPACITY
##          3.225925          1.219490          1.624007
```

We've seen most of the different models where we find possible high correlations between crimes, how some crime types might be confounding variable, and which other variables might affect r^2 value. However, since our goal is to learn the effect of crime on demand, we will include highly correlated variables back to the model. We will also remove EDU, since it might negatively impact the importance of other crime variables.

```
lm_bike_4 <- lm(trips ~ ASSAULT+ROBBERY+BURGLARY+THEFT+CRIMINAL_TRESPASS+NARCOTICS+HOMICIDE+STALKING+BA
summary(lm_bike_4)
```

```
##
## Call:
## lm(formula = trips ~ ASSAULT + ROBBERY + BURGLARY + THEFT + CRIMINAL_TRESPASS +
##     NARCOTICS + HOMICIDE + STALKING + BATTERY + DECEPTIVE_PRACTICE,
```

```
##     data = bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70777 -0.33371  0.02286  0.39757  1.21788
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.87923    0.45303  15.185  < 2e-16 ***
## ASSAULT            -1.13878    0.14926  -7.629 3.44e-13 ***
## ROBBERY            -0.28920    0.10987  -2.632  0.00894 **
## BURGLARY           -0.21563    0.06592  -3.271  0.00120 **
## THEFT               0.88286    0.13566   6.508 3.36e-10 ***
## CRIMINAL_TRESPASS   0.09237    0.09937   0.930  0.35335
## NARCOTICS           0.05704    0.08230   0.693  0.48879
## HOMICIDE           -0.10349    0.06280  -1.648  0.10044
## STALKING            0.15476    0.06448   2.400  0.01703 *
## BATTERY             0.36575    0.18591   1.967  0.05010 .
## DECEPTIVE_PRACTICE  0.31041    0.12694   2.445  0.01506 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5662 on 289 degrees of freedom
## Multiple R-squared:  0.6635, Adjusted R-squared:  0.6518
## F-statistic: 56.98 on 10 and 289 DF,  p-value: < 2.2e-16
```

We first estimate our regression model.

Hypothesis Null Hypothesis H_0: The crime variable has no effect on demand ($B_i = 0$) Alternative Hypothesis H_1: The crime variable has an effect on demand ($B_i \neq 0$)

From the Regression output and p-value, we can draw the following conclusion: Assault, Robbery, Burglary, Theft, Stalking, and Deceptive_Practice are signicant variables with p value < 0.05. Specficially, Assault, Bulglary, and Theft has p-value < 0.01. Battery will pass the hypothesis test at a significance level at the 0.1 level, meaning they have certain effect. Criminal Trespass, Narcotics, Homicide are not signicant variables.

From this hypothesis test, we can first notice that Drug and Weapon related crimes (Homicide) are not related to demand of the bikes. Both Burglary and Criminal Trespass involve unauthorized entry, but Burglary involves intent to commit a crime, while criminal trespass occurs when a person has no intent to commit a crime inside. There are two common points between these three kinds of crime: They are more likely to happen in private place, and they are not property crimes or impact. Although Burglary happens more likely in private place, but they will likely to involve property crime and has the intent to commit it. Violent crimes that are likely to happen in public place such as Assault, Robbery, Stalking, and Battery will negatively impact demand, as people rent bikes will use them in public area and feel dangerous by these violent crimes. Property Crimes that are likely to happen with the criminals have the intention to commit crime will also likely to negatively impact demand, as renters will feel their property unsecure. Thus, we can develop the following theorem:

Theorem: Violent crimes that occur in public spaces or involve criminal intent towards property will negatively impact bike rental demand, while crimes that occur in private spaces without intent to harm property will have minimal influence on demand.

Violent public crimes (Assault, Robbery, Battery, Stalking) reduce demand by creating perceptions of public danger.

All property crimes with criminal intent (Theft, Burglary) reduce demand by creating concerns about security and loss.

Other Private crimes or context-specific crimes (e.g., Homicide, Narcotics, Criminal Trespass) have little or no effect on demand, as they do not directly affect the public areas where bike rentals operate.

$ActualCrimeStatisticsvsPerceptionsofCrime$ Actual crime data from the Chicago Police Department measures real events, which are objective but may not reflect public perceptions or fears.

Perception-based fears might have a larger impact on demand than actual statistics. For example: If there's widespread media coverage of a single assault event in a bike-friendly area, it could have a larger impact on rentals than if statistics show that such events are rare.

In contrast, if narcotics offenses or homicide occur far from public bike routes, they might show no effect on rental demand despite their prevalence in crime data.