Statistical Modeling Course

Collinearity Lab

This lab focuses on the *collinearity* problem. Perform the following commands in `R` . The last line corresponds to creating a linear model in which `y` is a function of `x1` and `x2`.

```
set.seed(1)
x1 = runif(100)
x2 = 0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
df = tibble(y, x1, x2)
```

Problem 1

What is the correlation between `x1` and `x2`? What is the variance inflation factor? How about the condition number of $X^T X$?

```
# Correlation
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
# Variance Inflation Factor
model1 <- lm(y~x1+x2)
VIF(model1)
```

```
##       x1       x2
## 3.304993 3.304993
```

```
# Condition number
kappa(model1)
```

```
## [1] 13.28875
```

Problem 2

Using this data, fit a least squares regression to predict `y` using `x1` and `x2`. How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

```r
summary(fit <- lm(y~., df))
```

```
##
## Call:
## lm(formula = y ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

Answer: The true values are $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$, $\hat{\beta}_1 = 1.43$ is too low while $\hat{\beta}_2 = 1.00$ is too high. We can reject $H_0 : \beta_1 = 0$ but not $H_0 : \beta_2 = 0$

Problem 3

Now fit a least squares regression to predict `y` using only `x1` . Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```r
summary(lm(y~x1, df))
```

```
##
## Call:
## lm(formula = y ~ x1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

Answer: The estimate for β_1 is now closer to the true value and we can reject $H_0 : \beta_1 = 0$.

Problem 4

Now fit a least squares regression to predict `y` using only `x2` . Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```r
summary(lm(y~x2, df))
```

```
##
## Call:
## lm(formula = y ~ x2, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

Answer: The estimate for β_1 is much higher than the true value. In this case we can reject $H_0 : \beta_1 = 0$.

Problem 5

Do the results obtained in Problem 2 and 4 contradict each other? Explain your answer. *Answer: No, because the two variables are correlated and x1 has a stronger effect if we include x1 in the model the effect of x2 is not significant. However if we include just x2 without x1 there is a large significant effect.*