STAT 420: Quiz Assignment

Nico Kienawan, nicok2

4/30/2020

Collinearity

Consider the following model,

$$Y = 4 + 3X_1 + 2X_2 + X_3 + \epsilon,$$

where $\epsilon \sim N(\mu = 0, \sigma = 3)$.
Simulate a sample size of 100 observations from this model.

```r
n = 100
set.seed(910)
x1 = runif(n, 0 , 10)
x2 = runif(n, 0 , 10)
y = function(x1, x2, x3) {4 + 3 * x1 + 2 * x2 + x3 + rnorm(n, 0 , 3)}
```

(a) Let $X_3 = 6X_1 + 5X_2$, save this as `x3_a`. Using `lm()`, fit a simple linear model and save the model as `model_a`. Then, print the `summary()` of the model. Describe anything unusual from the `summary()`.

```r
x3_a = 6 * x1 + 5 * x2
y_a = y(x1, x2, x3_a)
model_a = lm(y_a ~ x1 + x2 + x3_a)
summary(model_a)
```

```
##
## Call:
## lm(formula = y_a ~ x1 + x2 + x3_a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7237 -1.5196 -0.2401  1.5426  6.1451
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.77144    0.62593   7.623 1.68e-11 ***
## x1           9.03216    0.08994 100.428  < 2e-16 ***
## x2           6.83283    0.08718  78.372  < 2e-16 ***
## x3_a              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.483 on 97 degrees of freedom
## Multiple R-squared:  0.9948, Adjusted R-squared:  0.9947
## F-statistic:  9232 on 2 and 97 DF,  p-value: < 2.2e-16
```

The values in the `x3` row are all NAs.

(b) Now, let $X_3 = 6X_1 + 5X_2 + \epsilon$ where $\epsilon \sim N(\mu = 0, \sigma = 0.1)$, save this as `x3_b`. Using `lm()`, fit a simple linear model and save the model as `model_b`. Then, print the `summary()` of the model. Report the R-squared and the p-value of each predictors. Do you see any differences from the `summary()` in part **(a)**? Why are they different?

```
x3_b = 6 * x1 + 5 * x2 + rnorm(n, 0, 0.1)
y_b = y(x1, x2, x3_b)
model_b = lm(y_b ~ x1 + x2 + x3_b)
summary(model_b)
```

```
##
## Call:
## lm(formula = y_b ~ x1 + x2 + x3_b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2818 -1.7153  0.2859  1.9853  5.7144
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0137     0.7025   7.137 1.81e-10 ***
## x1           -6.5509    16.4803  -0.398    0.692
## x2           -5.9017    13.6945  -0.431    0.667
## x3_b          2.5715     2.7442   0.937    0.351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.78 on 96 degrees of freedom
## Multiple R-squared:  0.9935, Adjusted R-squared:  0.9933
## F-statistic:  4874 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
summary(model_b)$r.squared
```

```
## [1] 0.9934775
```

```
summary(model_b)$coefficient[2:4,4]
```

```
##        x1        x2      x3_b
## 0.6918804 0.6674712 0.3510628
```

The values in the `x3` row are not NAs anymore.
They are different because the `x3` in `model_a` is exactly linearly dependent to `x1` and `x2`. While in `model_b`, the `x3` is still linearly dependent to `x1` and `x2` but not exactly.

(c) Notice that the R-squared in `model_b` is very close to 1. However, none of the predictors are significant. Why do you think this happens?
This happens because the predictors in `model_b` is highly correlated (`x3` is very linearly dependent to `x1` and `x2` with the σ of ϵ only 0.1) which reduces the significance of `x1` and `x2`.

(d) Suppose `X = cbind(1, x1, x2, x3_b)`. Find the eigenvalues of $X^T X$. Report the smallest eigenvalue. What do you think is the relationship between this eigenvalue and the significance of the t-test above? (Remember: $(X^T X)^{-1} = QD^{-1}Q^T$ where D is a diagonal matrix with the eigenvalues of $X^T X$ and variance of $\hat{\beta}$ is $\hat{\sigma}^2(X^T X)^{-1}$)

```r
x = cbind(1, x1, x2, x3_b)
eigenval = eigen(t(x) %*% x)$values
eigenval
```

```
## [1] 3.439976e+05 7.194713e+02 1.571251e+01 1.655574e-02
```

```r
eigenval[which.min(eigenval)]
```

```
## [1] 0.01655574
```

Since an eigenvalue of X^TX is very small, $(X^TX)^{-1}$ will be very large because $(X^TX)^{-1} = QD^{-1}Q^T$.
If $(X^TX)^{-1}$ is very large, the variance of $\hat{\beta}$ will be very large too because $Var(\hat{\beta}) = \hat{\sigma}^2(X^TX)^{-1}$ and $\hat{\sigma}^2$ is around 1.
Since the variance of $\hat{\beta}$ is very large, the confidence interval of the test will also be large, which makes the t-test above not significant.

(e) Find the Variance Inflation Factor (VIF) of `model_b`. Do you find any "problematic" predictors? What does it mean?

```r
library(car)
```

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

```r
vif(model_b)
```

```
##       x1       x2     x3_b
## 27206.09 19990.38 52980.63
```

Yes, all of the predictors are "problematic" because the VIFs are very large.
It means there are correlations between the predictors.

Appendix

- The Collinearity problem used the Multicollinearity_SLIDES powerpoint as a reference.