Lab-8 Multicollinearity

PrabhTalwar

2022-12-09

```
library(readr)
ftccigar <- read_csv("ftccigar.csv")
ftccigar <- ftccigar[,-1]
View(ftccigar)
```

(a) Fit the model to the data. Is there evidence that tar content x1, is useful for predicting carbon monoxide content, y?

```
ftc_model_1 <- lm(CO ~ TAR, data = ftccigar)
summary(ftc_model_1)
```

```
##
## Call:
## lm(formula = CO ~ TAR, data = ftccigar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1124 -0.7167 -0.3754  1.0091  2.5450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74328    0.67521   4.063 0.000481 ***
## TAR          0.80098    0.05032  15.918 6.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 23 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.9132
## F-statistic: 253.4 on 1 and 23 DF,  p-value: 6.552e-14
```

The TAR content, is useful for predicting carbon monoxide content as the Adjusted R-squared is also very high which is 0.9132.

(b) Fit the model to the data. Is there evidence that nicotine content x2, is useful for predicting carbon monoxide content, y?

```
ftc_model_2 <- lm(CO ~ NICOTINE, data = ftccigar)
summary(ftc_model_2)
```

```
##
## Call:
## lm(formula = CO ~ NICOTINE, data = ftccigar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3273 -1.2228  0.2304  1.2700  3.9357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6647     0.9936   1.675    0.107
## NICOTINE     12.3954     1.0542  11.759 3.31e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.828 on 23 degrees of freedom
## Multiple R-squared:  0.8574, Adjusted R-squared:  0.8512
## F-statistic: 138.3 on 1 and 23 DF,  p-value: 3.312e-11
```

The NICOTINE content, is useful for predicting carbon monoxide content as the Adjusted R-squared is also high which is 0.8512

(c) Fit the model to the data. Is there evidence that weight content x3, is useful for predicting carbon monoxide content, y?

```
ftc_model_3 <- lm(CO ~ WEIGHT, data = ftccigar)
summary(ftc_model_3)
```

```
##
## Call:
## lm(formula = CO ~ WEIGHT, data = ftccigar)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.524 -2.533  0.622  2.842  7.268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11.795      9.722  -1.213   0.2373
## WEIGHT        25.068      9.980   2.512   0.0195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.289 on 23 degrees of freedom
## Multiple R-squared:  0.2153, Adjusted R-squared:  0.1811
## F-statistic: 6.309 on 1 and 23 DF,  p-value: 0.01948
```

The WEIGHT content, is not that useful for predicting carbon monoxide content as the Adjusted R-squared is also very low which is 0.1811

```
coef(ftc_model_1)
```

```
## (Intercept)         TAR
##    2.743278    0.800976
```

```
coef(ftc_model_2)
```

```
## (Intercept)    NICOTINE
##    1.664666   12.395406
```

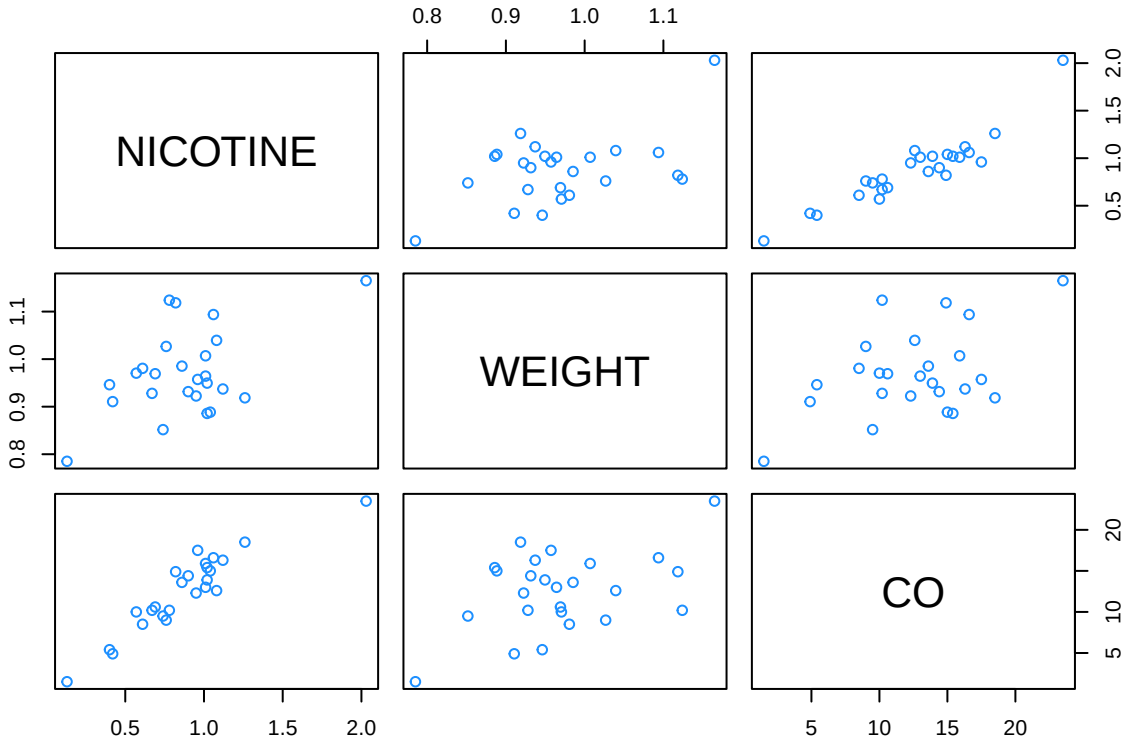
```
coef(ftc_model_3)
```

```
## (Intercept)      WEIGHT
##   -11.79527    25.06820
```

However, the estimated coefficients are wildly different for all the three models.

(d) Fit the model as we did in the notes for all three variables, Compare the signs of beta1, beta2 and beta3 in the models of parts (a),(b) and (c) to the signs of the betas in the model with all three betas. Is the fact that the betas change dramatically when the independent variables are removed from the model an indication of a serious multicollinearity problem?

```
pairs(ftccigar[,-1], col = "dodgerblue")
```




```
round(cor(ftccigar[,-1]),2)
```

```
##          NICOTINE WEIGHT   CO
## NICOTINE     1.00   0.50 0.93
## WEIGHT       0.50   1.00 0.46
## CO           0.93   0.46 1.00
```

```
ftc_model <- lm(CO ~ ., data = ftccigar)
summary(ftc_model)
```

```
##
## Call:
## lm(formula = CO ~ ., data = ftccigar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89261 -0.78269  0.00428  0.92891  2.45082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2022     3.4618   0.925 0.365464
## TAR           0.9626     0.2422   3.974 0.000692 ***
## NICOTINE     -2.6317     3.9006  -0.675 0.507234
## WEIGHT       -0.1305     3.8853  -0.034 0.973527
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 on 21 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.907
## F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

One of the first things we notice that the F-test for the regression tells us that the regression is significant, however each individual predictor is not. Another interesting result is the opposite signs of the coefficients for NICOTINE and WEIGHT. This is a case of high correlation.

Using the VIF, as it helps to understand how collinearity affects our regression estimates.

```
faraway::vif(ftc_model)
```

```
##       TAR  NICOTINE    WEIGHT
## 21.630706 21.899917  1.333859
```

If VIF is greater than 5 causes problems. Here, we see huge multicollinearity issue as TAR and NICOTINE have a VIF greater than 5.