Regression Analysis 011224 James Ward

Final Project (PASS/ACE/Certificate)

Problem 5.13

This exercise is an open-ended challenge to fit a multiple linear regression model to some data on restaurants. Try to follow the model building guidelines in Section 5.3 as best you can, and strive to come up with a "good" model (for this application, a "good" model should have an R-squared value of approximately 0.94 and a regression standard error, s, of approximately 10). You could potentially spend many hours on this exercise, but it should be possible to come up with a decent model within an hour or so; if you find yourself spending much more time than this, chances are you're on the wrong track or you're working too hard!

The following problem provides a challenging dataset that you can use to practice multiple linear regression model building. You've been asked to find out how profits for 120 restaurants in a particular restaurant chain are affected by certain characteristics of the restaurants. You would like to build a regression model for predicting Profit = annual profits (in thousands of dollars) from five potential predictor variables:

Cov = number of covers or customers served (in thousands)
Fco = food costs (in thousands of dollars)
Oco = overhead costs (in thousands of dollars)
Lco = labor costs (in thousands of dollars)
Region = geographical location (Mountain, Southwest, or Northwest)

Note that region is a qualitative (categorical) variable with three levels; the **RESTAURANT** data file contains two indicator variables to code the information in region: D_{Sw} = 1 for Southwest, 0 otherwise, and D_{Nw} = 1 for Northwest, 0 otherwise. Thus, the Mountain region is the reference level with 0 for both D_{Sw} and D_{Nw}. Build a suitable regression model and investigate the role of each of the predictors in the model through the use of predictor effect plots. You may want to consider the following topics in doing so:

- models with both quantitative and qualitative variables;
- polynomial transformations;
- interactions;
- comparing nested models.

You may use the following for terms in your model:

- Cov, Fco, Oco, Lco, DSw, DNw;

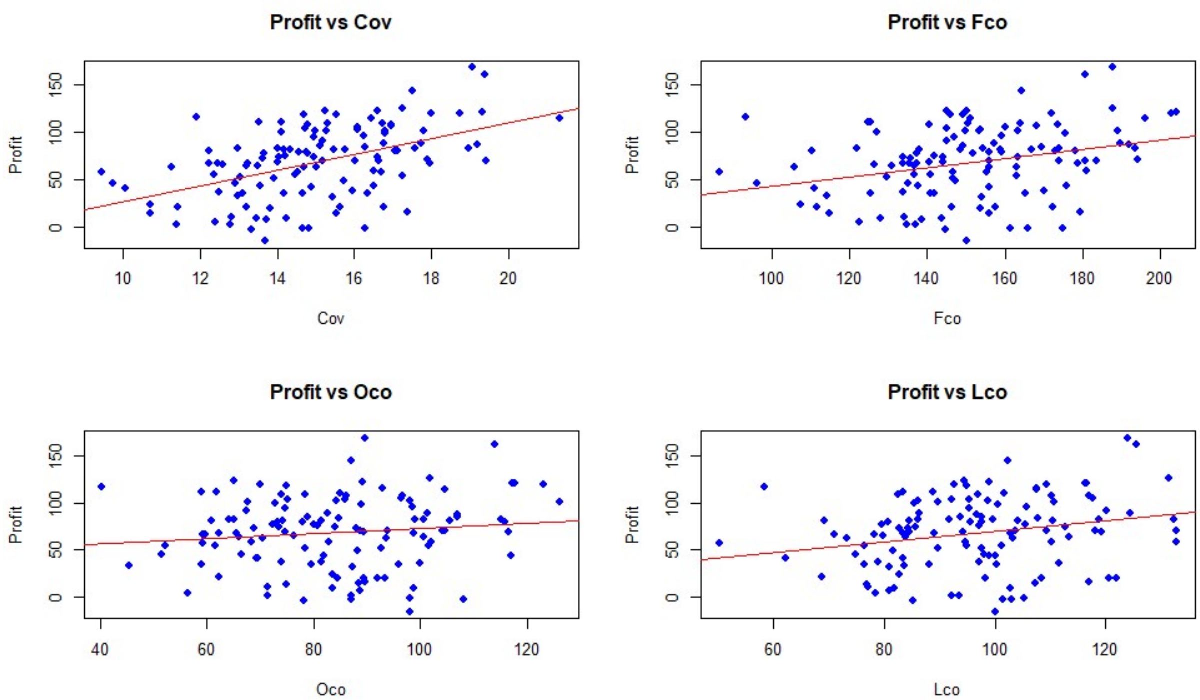
- interactions between each of the quantitative predictors and the indicator variables, such as DSwCov, DNwCov, etc.;
- quadratic terms, such as Cov2 (do not use terms like DSw2, however!);

use Profit as the response variable [i.e., do not use loge(Profit) or any other transformation].

First, I would like to look at what the linear model would look like:

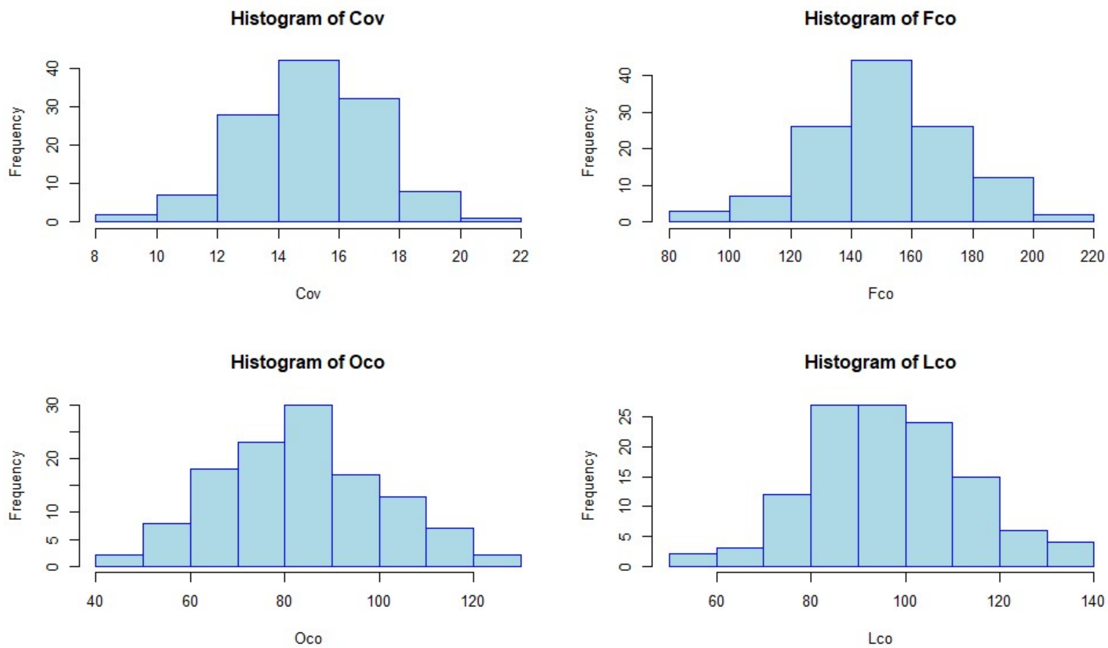
Profit = B0 + B1(Cov) + B2(Fco) + B3(Oco) + B4(Lco) + B5(DSw) + b6(Dnw)

Basic Normality Assumptions:

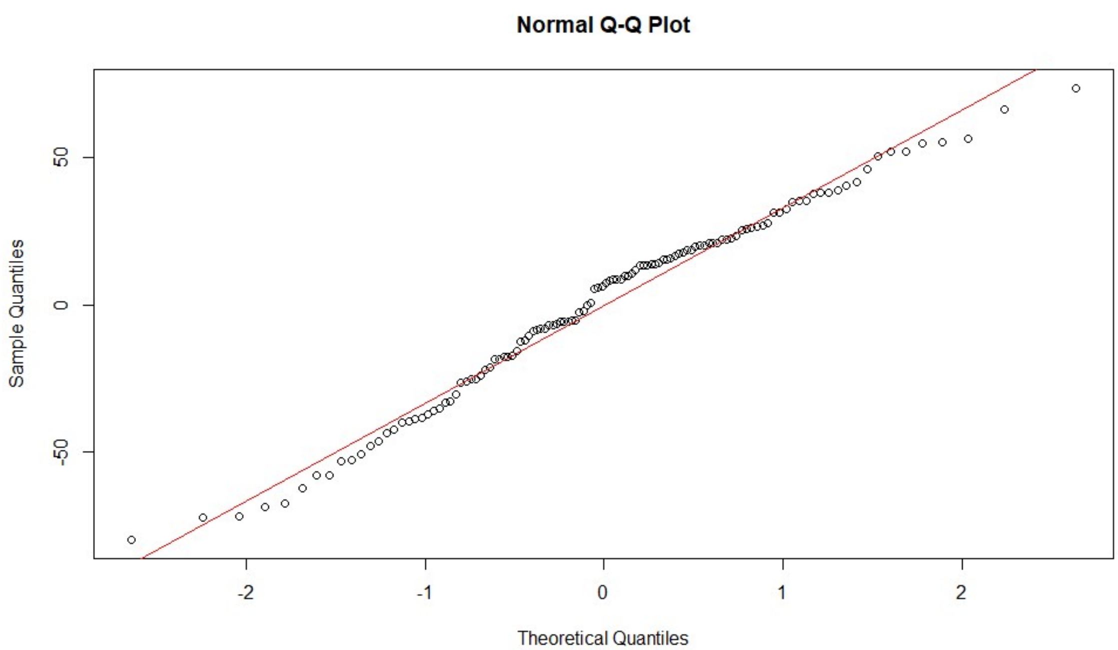


The scatter plots with least squares lines of the predictors look normally distributed.

The constant variance looks good.



The histograms of the predictors look normally distributed.



The QQ plots follow the linear regression line.

Fit the linear regression

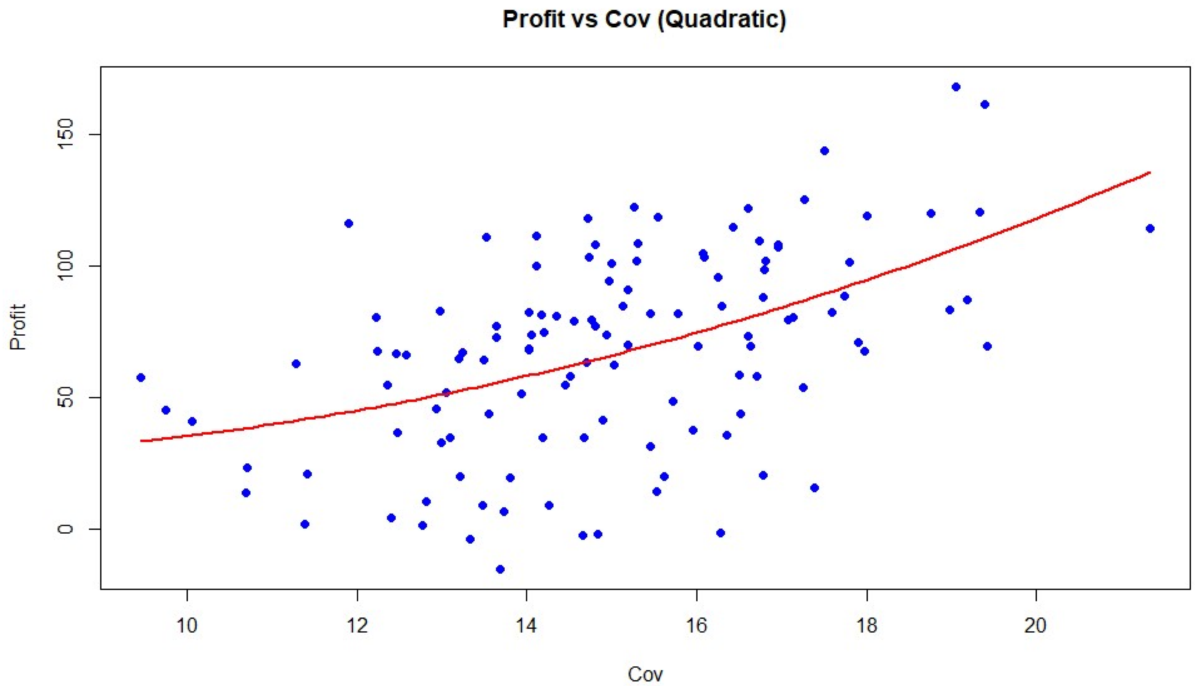
```
summary(initial_model)

Call:
lm(formula = Profit ~ Cov + Fco + Oco + Lco + DSw + DNw, data = restaurant)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.502  -7.531  -2.031   7.842  39.232 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -55.8406     7.8369  -7.125 1.04e-10 ***
Cov          25.3628     1.1855  21.394  < 2e-16 ***
Fco          -0.9001     0.1427  -6.307 5.71e-09 ***
Oco          -0.5438     0.1041  -5.226 8.02e-07 ***
Lco          -0.6271     0.1171  -5.357 4.51e-07 ***
DSw          10.8023     2.6141   4.132 6.92e-05 ***
DNw         -58.2642     2.7243 -21.387  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.85 on 113 degrees of freedom
Multiple R-squared:  0.9078,    Adjusted R-squared:  0.9029 
F-statistic: 185.4 on 6 and 113 DF,  p-value: < 2.2e-16
```

Each of the predictors has a very low p-value so they are strong predictors. The indication is to reject a null hypothesis that the predictors are statistically insignificant.

The Residual Standard Error is on average 11.85 thousand dollars away from the true regression. The desired amount is 10.

The R squared value is .9078. The desired value is .94.

Lets move on to the quadratic equation.



summary(quadratic_model)

```
Call:
lm(formula = Profit ~ Cov + Fco + Oco + Lco + DSw + DNw + I(Cov^2),
    data = restaurant)

Residuals:
    Min      1Q  Median      3Q     Max
-27.963  -6.916  -2.028   8.540  38.806

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.5292    36.0707  -0.569 0.570402
Cov          20.3670     5.1205   3.978 0.000124 ***
Fco          -0.8836     0.1437  -6.150 1.23e-08 ***
Oco          -0.5582     0.1051  -5.314 5.52e-07 ***
Lco          -0.6122     0.1180  -5.189 9.52e-07 ***
DSw          10.6972     2.6161   4.089 8.19e-05 ***
DNw         -58.2295     2.7245 -21.373  < 2e-16 ***
I(Cov^2)      0.1607     0.1602   1.003 0.318068
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.85 on 112 degrees of freedom
Multiple R-squared:  0.9086,    Adjusted R-squared:  0.9029
F-statistic: 159.1 on 7 and 112 DF,  p-value: < 2.2e-16
```

The scatterplot shows a normal distribution of the data points along the quadratic regression line.

The Residual Standard Error is 11.85 which is still greater than 10.

```
The Multiple R-Squared is .9086 which is still lower than the desired .94.
```

Polynomial transformation

```
summary(quadratic_model2)

Call:
lm(formula = Profit ~ Cov + I(Cov^2) + I(Cov^3) + Fco + I(Fco^2) +
    Oco + Lco + DSw + DNw, data = restaurant)

Residuals:
    Min      1Q  Median      3Q     Max
-30.642  -6.744  -1.510   7.776  33.228

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.234603 158.072362   0.362 0.717988
Cov          21.538224  35.172216   0.612 0.541560
I(Cov^2)      0.573932   2.272528   0.253 0.801086
I(Cov^3)     -0.019425   0.048453  -0.401 0.689265
Fco          -2.524637   0.731718  -3.450 0.000795 ***
I(Fco^2)      0.005440   0.002366   2.299 0.023384 *
Oco          -0.576439   0.105355  -5.471 2.83e-07 ***
Lco          -0.603311   0.115719  -5.214 8.76e-07 ***
DSw          12.140295   2.637897   4.602 1.13e-05 ***
DNw         -56.568566   2.745055 -20.607  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.61 on 110 degrees of freedom
Multiple R-squared:  0.9139,    Adjusted R-squared:  0.9068
F-statistic: 129.7 on 9 and 110 DF,  p-value: < 2.2e-16
```

The cov and the polynomials attempt to model a non linear relationship between cov and profit, however, none of these terms are statistically significant p>.05.

The residual Standard Error improved slightly to 11.61.

The multiple R-squared improved slightly to .9139

Interactions

My first interaction is to add a term between ‘Cov’ and ‘DSw’. Hoping to find an added effect of Cov and Profit in the Southwest Region.

```
Call:
lm(formula = Profit ~ Cov + Fco + Oco + Lco + DSw + DNw + I(Cov^2) +
    Cov:DSw, data = restaurant)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-25.980  -7.779  -0.974   6.887  31.924 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   33.1218    38.1252   0.869  0.38685    
Cov           13.7267     5.2948   2.592  0.01081 *  
Fco           -0.8046     0.1396  -5.762 7.54e-08 ***
Oco           -0.5591     0.1006  -5.556 1.92e-07 ***
Lco           -0.6420     0.1134  -5.663 1.19e-07 ***
DSw          -41.7240    15.9481  -2.616  0.01013 *  
DNw          -57.5898     2.6167 -22.009  < 2e-16 ***
I(Cov^2)       0.3213     0.1609   1.998  0.04821 *  
Cov:DSw        3.5671     1.0717   3.328  0.00119 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.35 on 111 degrees of freedom
Multiple R-squared:  0.9169,    Adjusted R-squared:  0.9109 
F-statistic: 153.1 on 8 and 111 DF,  p-value: < 2.2e-16
```

Ineratction of Quadratic: Cov and DSw

```
Call:
lm(formula = Profit ~ Cov * DSw + I(Cov^2) * DSw + Fco + Oco + 
    Lco + DNw, data = restaurant)

Residuals:
     Min       1Q   Median       3Q      Max 
-25.9780  -8.1536  -0.7682   7.0907  31.5736 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     5.5942    47.3072   0.118  0.90608    
Cov            17.5217     6.5532   2.674  0.00864 ** 
DSw            29.1484    73.8324   0.395  0.69376    
I(Cov^2)        0.2015     0.2019   0.998  0.32035    
Fco            -0.7974     0.1399  -5.701 1.01e-07 ***
Oco            -0.5722     0.1015  -5.636 1.36e-07 ***
Lco            -0.6612     0.1151  -5.746 8.24e-08 ***
DNw           -57.6851     2.6189 -22.027  < 2e-16 ***
Cov:DSw        -6.2465    10.0394  -0.622  0.53510    
DSw:I(Cov^2)    0.3326     0.3383   0.983  0.32770    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.35 on 110 degrees of freedom
Multiple R-squared:  0.9176,    Adjusted R-squared:  0.9109 
F-statistic: 136.2 on 9 and 110 DF,  p-value: < 2.2e-16
```

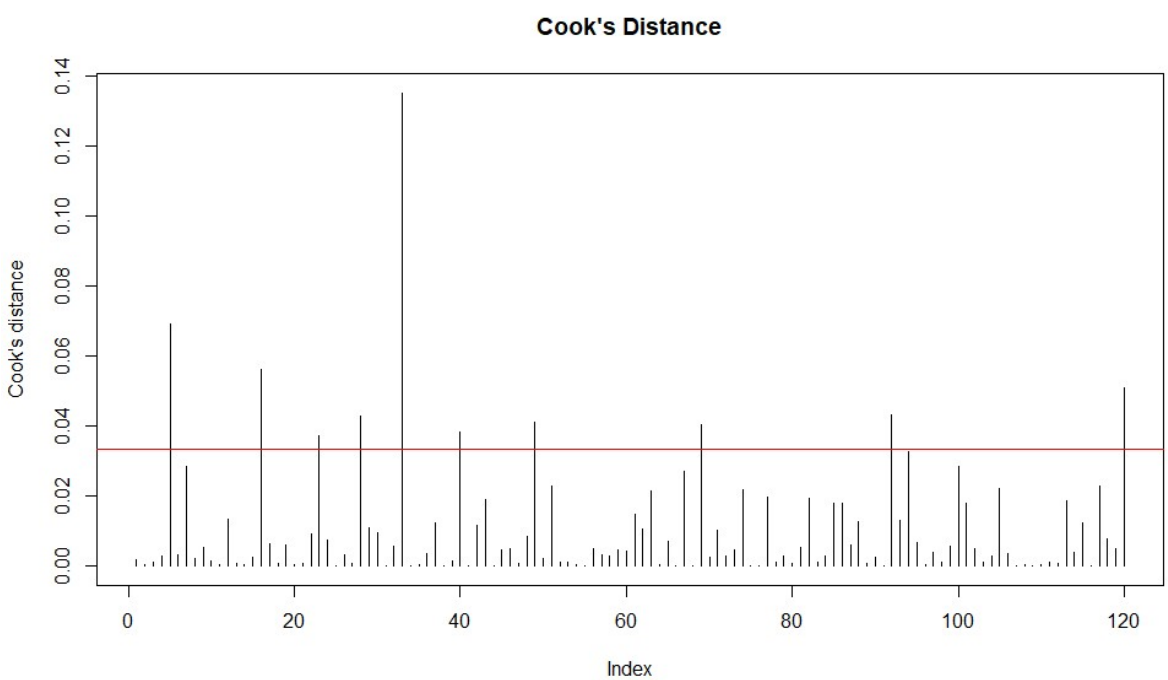
Interaction of Cov:DSw and Cov:DNw

```
Call:
lm(formula = Profit ~ Cov + I(Cov^2) + Fco + I(Fco^2) + Oco + 
```

```
Lco + DSw + DNw + Cov:DSw + Cov:DNw, data = restaurant)

                    Residuals:
    Min      1Q  Median      3Q     Max
-27.788  -6.258  -1.250   6.742  29.391

                    Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
 (Intercept)  47.246129  36.977093   1.278  0.20407
Cov          21.809738   8.023811   2.718  0.00764 **
 I(Cov^2)      0.121297   0.264102   0.459  0.64695
Fco          -1.943658   0.677895  -2.867  0.00497 **
 I(Fco^2)      0.003617   0.002203   1.642  0.10353
Oco          -0.591655   0.097490  -6.069 1.91e-08 ***
Lco          -0.637513   0.109264  -5.835 5.61e-08 ***
 DSw         -19.550184  17.116180  -1.142  0.25587
 DNw         -11.192022  19.578285  -0.572  0.56873
 Cov:DSw       2.143044   1.149231   1.865  0.06490 .
 Cov:DNw      -2.992750   1.289435  -2.321  0.02215 *
                                    ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 10.93 on 109 degrees of freedom
 Multiple R-squared:  0.9244,    Adjusted R-squared:  0.9175
  F-statistic: 133.3 on 10 and 109 DF,  p-value: < 2.2e-16
```

Better!!!

<u>Now I will remove the outliers using Cook's Distance.</u>



Observations with high Cook's distance: 5 16 23 28 33 40 49 69 92 120

After removing the observations:

```
# Summary of the model after removing outliers
> summary(simplified_model_clean)

Call:
lm(formula = Profit ~ Cov + Fco + I(Fco^2) + Oco + Lco + DSw +
    DNw + Cov:DSw + Cov:DNw, data = restaurant_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-20.243  -5.816  -1.572   6.282  23.907

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.504446  36.948872  -0.284 0.776771
Cov          27.403907   1.394349  19.654  < 2e-16 ***
Fco          -1.808400   0.462135  -3.913 0.000166 ***
I(Fco^2)      0.002846   0.001539   1.849 0.067344 .
Oco          -0.612122   0.090044  -6.798 7.81e-10 ***
Lco          -0.628968   0.098982  -6.354 6.26e-09 ***
DSw           2.058425  18.194477   0.113 0.910151
DNw          -1.375492  17.273288  -0.080 0.936690
Cov:DSw       0.573295   1.224446   0.468 0.640655
Cov:DNw      -3.666490   1.133294  -3.235 0.001648 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.479 on 100 degrees of freedom
Multiple R-squared:  0.9406,    Adjusted R-squared:  0.9352
F-statistic: 175.9 on 9 and 100 DF,  p-value: < 2.2e-16
```

I have obtained the desired threshold. I will stop now.

Code pasted below:

```
# Calculate Cook's distance
> cooks_distances <- cooks.distance(simplified_model)
>
> # Plot Cook's distance
> plot(cooks_distances, ylab="Cook's distance", type='h', main="Cook's Distance")
> abline(h=4/length(cooks_distances), col="red")  # Threshold line
>
> # Identify observations with Cook's distance greater than the threshold
> threshold <- 4/length(cooks_distances)
> influential_points <- which(cooks_distances > threshold)
>
> cat("Observations with high Cook's distance:", influential_points, "\n")
Observations with high Cook's distance: 5 16 23 28 33 40 49 69 92 120
>
> # Removing observations identified as influential
> restaurant_clean <- restaurant[-influential_points, ]
>
> # Fit the model again with the cleaned data
> simplified_model_clean <- lm(Profit ~ Cov + Fco + I(Fco^2) + Oco + Lco + DSw +
DNw + Cov:DSw + Cov:DNw, data=restaurant_clean)
```

```
>
> # Summary of the model after removing outliers
> summary(simplified_model_clean)
```