Class_Activity#7

Sunny Lee

2021-02-24

Interaction and Terms for Curvature

Use the 4-8Cyl data set for this class activity.

```
cyl <- read.csv("C3 4_8Cyl.csv")
attach(cyl)
```

1) Use the 4-8Cyl data set to calculate the two regression equations

```
m1 <- lm(Price~Mileage+Cyl)
summary(m1)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Cyl)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16815  -8641   1410   5369  27978
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.535e+04  2.207e+03   6.955 3.29e-11 ***
## Mileage     -2.001e-01  6.779e-02  -2.952  0.00347 **
## Cyl          3.443e+03  3.044e+02  11.312  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8927 on 241 degrees of freedom
## Multiple R-squared:  0.3639, Adjusted R-squared:  0.3587
## F-statistic: 68.95 on 2 and 241 DF,  p-value: < 2.2e-16
```

```
#interaction terms
m2 <- lm(Price~Mileage*Cyl)
summary(m2)
```

```
##
## Call:
## lm(formula = Price ~ Mileage * Cyl)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17101  -8824   1369   5365  24755
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4533.02363 4368.42824   1.038  0.30046
## Mileage        0.34006    0.20066   1.695  0.09142 .
## Cyl         5430.70175  758.07201   7.164 9.58e-12 ***
## Mileage:Cyl   -0.09953    0.03486  -2.855  0.00468 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8797 on 240 degrees of freedom
## Multiple R-squared:  0.3848, Adjusted R-squared:  0.3771
## F-statistic: 50.05 on 3 and 240 DF,  p-value: < 2.2e-16
```

$$Price = \beta_0 + \beta_1(Mileage) + \beta_2 * Cyl) + \epsilon,$$

$$Price = \beta_0 + \beta_1(Mileage) + \beta_2 * Cyl) + \beta_3(Mileage * Cyl) + \epsilon$$

a) Does the R^2_{adj} value increase when the interaction term is added? Based on the change in R^2_{adj}, should the interaction term be included in the model?

The R^2_{adj} value increase is about 2 percent when we add our interaction term is added. We conclude that this is not a large enough change to conclude the itneraction term is necessary.

b) For both models, calculate the estimate price of a four cyliner car when Mileage=10,000.

For the first model: $15350 - 0.2001(10000) + 3443(4) = 27121$ and for the second model: $4533.02363 + 0.34006(10000) + 5430.70175(4) - 0.09953(10000)(4) = 25675.23063$

c) Assume Mileage=10,000, for both models explain how increasing from four to eight cylinder will impact the estimated price.

Changing our calculations from 4 to 8 cylinders: our first model: $15350 - 0.2001(10000) + 3443(8) = 40893$ and our second: $4533.02363 + 0.34006(10000) + 5430.70175(8) - 0.09953(10000)(8) = 43416.83763$ and thus, we observe a $40893 - 27121 = 13772$ dollar difference going from 4 to 8 cylinders in the first model and a $43416.83763 - 25675.23063 = 17741.607$ dollar difference going from 4 to 8 cylinders in the second model.

d) conduct an extra sum of squares test to determine if the Mileagecyl interaction term is important to the model.

```
anova(m1, m2)
```

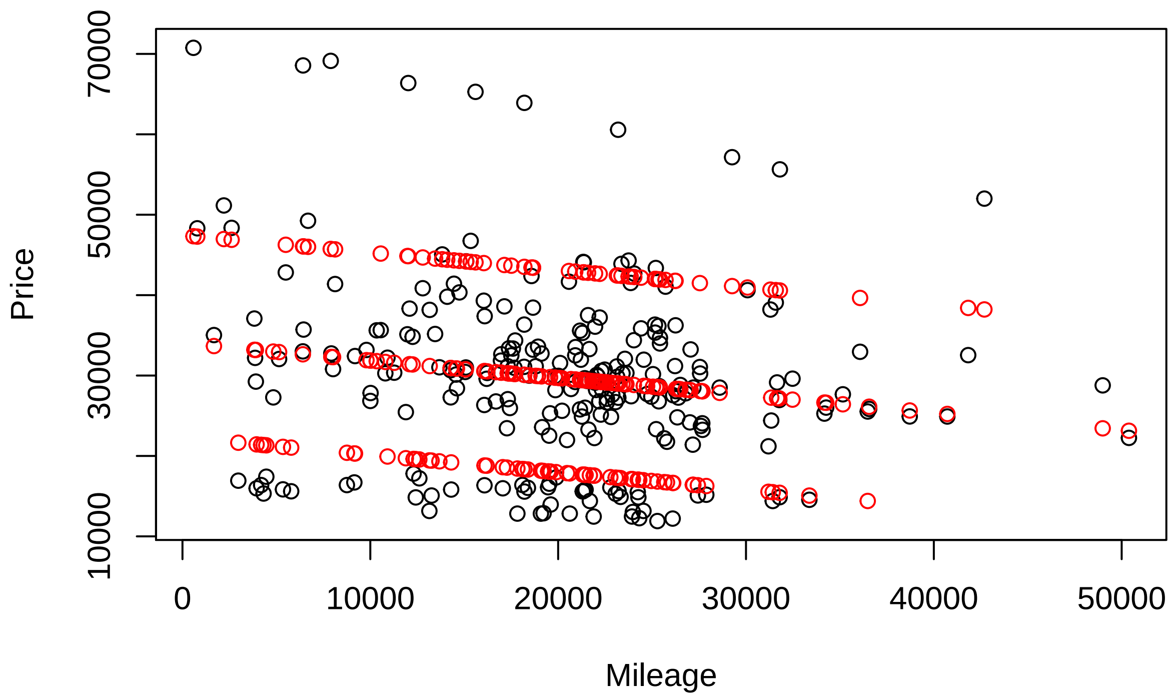
```
## Analysis of Variance Table
##
## Model 1: Price ~ Mileage + Cyl
## Model 2: Price ~ Mileage * Cyl
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    241 1.9205e+10
## 2    240 1.8574e+10  1 630762295 8.1502 0.004682 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running our extra sum of squares test, we see our p value is quite low and thus we would have to reject the null hypothesis. Since we reject the null hypothesis, we conclude that the second model significant and that we should use the model with the interaction term.

2) Use the 4-8Cyl data set to calculate the regression line $Price = \beta_0 + \beta_1(Mileage) + \beta_2(Cadilac) + \beta_3(SAAB)$. You will need to create indicator variables for Make before calculating the regression line.

a) Create a scatterplot with Mileage as the explanatory variable and Price as the response. Overlay a second graph with Mileage as the explanatory variable and $\hat{y}$ as the response. Notice that the predicted value (the $\hat{y}$ values) form two seperate lines. Do the parallel lines (no interaction model) look appropriate?

```r
Cadilac <- (Make == "Cadillac") * 1
SAAB <- (Make == "SAAB") * 1
m3 <- lm(Price~Mileage+Cadilac+SAAB)
m4 <- lm(Price~Mileage*SAAB + Mileage*Cadilac)
plot(Price~Mileage)
points(m3$fitted.values~Mileage, col = "red")
```



The three parallel lines are the result of our categorical variables Cadillac and SAAB. Our model depends on Cadillac and SAAB, however there are other makes in the data as well. Thus, one of our red lines will model when Cadillac is 1, one will model when SAAB is 1 and the last will model when neither are 1. Thus, the parallel lines look appropriate considering our model.

`b) Conduct one extra sum of squares test to determine if interaction term (MileageCadilac and MileageSA`

```r
summary(m3)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Cadilac + SAAB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11319.2  -4013.9   -984.7   1895.0  23430.5
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.229e+04  1.157e+03  19.268  < 2e-16 ***
## Mileage     -2.162e-01  4.658e-02  -4.641  5.7e-06 ***
## Cadilac      2.516e+04  1.074e+03  23.428  < 2e-16 ***
## SAAB         1.174e+04  9.304e+02  12.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6105 on 240 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.7001
## F-statistic:   190 on 3 and 240 DF,  p-value: < 2.2e-16
```

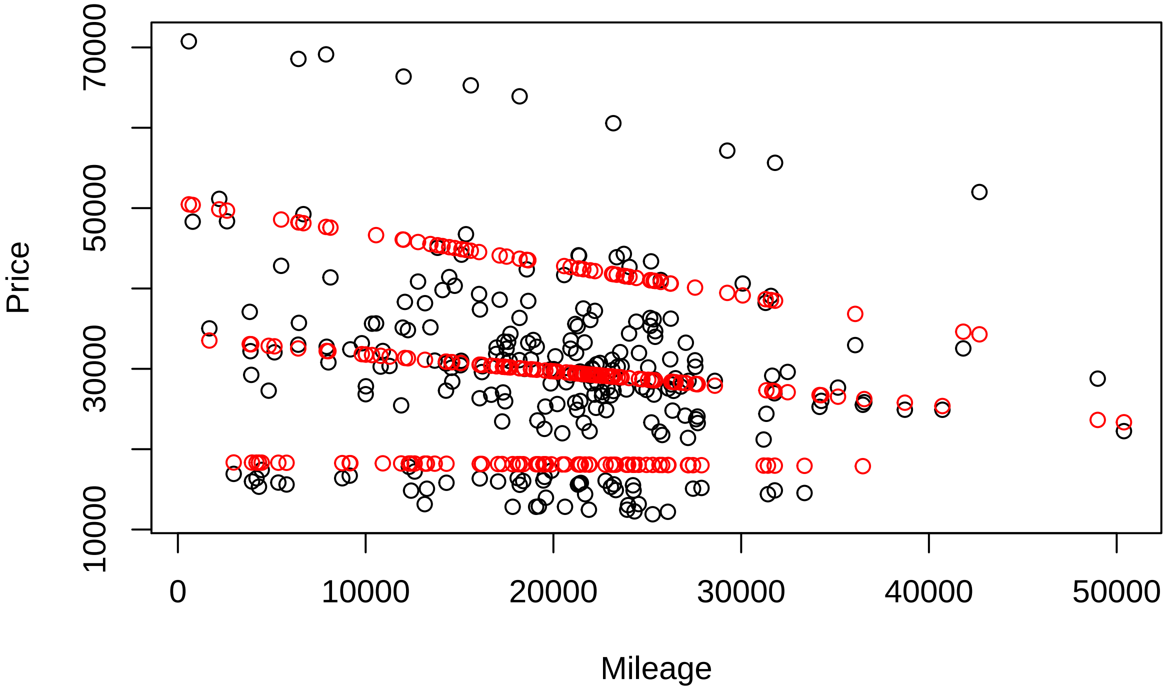
```
summary(m4)
```

```
##
## Call:
## lm(formula = Price ~ Mileage * SAAB + Mileage * Cadilac)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12502  -3626  -1001   1880  21471
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.839e+04  1.969e+03   9.338  < 2e-16 ***
## Mileage         -1.368e-02  9.510e-02  -0.144  0.88578
## SAAB             1.549e+04  2.488e+03   6.224 2.17e-09 ***
## Cadilac          3.231e+04  2.663e+03  12.134  < 2e-16 ***
## Mileage:SAAB    -1.952e-01  1.166e-01  -1.674  0.09535 .
## Mileage:Cadilac -3.705e-01  1.267e-01  -2.925  0.00377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6023 on 238 degrees of freedom
## Multiple R-squared:  0.7141, Adjusted R-squared:  0.7081
## F-statistic: 118.9 on 5 and 238 DF,  p-value: < 2.2e-16
```

```
anova(m3, m4)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Mileage + Cadilac + SAAB
## Model 2: Price ~ Mileage * SAAB + Mileage * Cadilac
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1    240 8944659031
## 2    238 8633414669  2 311244362 4.2901 0.01478 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(Price~Mileage)
points(m4$fitted.values~Mileage, col = "red")
```



From our anova table, we find that our p value for extra sum of squares is quite low, meaning we can reject the null hypothesis. Thus, we should be able to conclude that our interaction terms are significant in predicting the price. We can also see this is the case, as in our plot we get three non-parallel lines which would indicate that our interaction terms are affecting our model, as we had three parallel lines before.