Project 2

Multiple Regression Models for Car Prices

Carley Dziewicki

Febuary 28, 2020

When do we use prediction intervals over confidence interval

```
# Be sure to include command(s) to import your data here
library(readr)
OutbackData <- read_csv("~/Desktop/Stats2-Math360/OutbackData.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Price = col_double(),
##   Mileage = col_double(),
##   Age = col_double()
## )
```

```
#View(OutbackData)
OutbackData <- mutate(OutbackData, Price = Price/1000)
head(OutbackData)
```

```
## # A tibble: 6 x 4
##    Year Price Mileage   Age
##   <dbl> <dbl>   <dbl> <dbl>
## 1  2017  20.4   37568     3
## 2  2017  22.3   43372     3
## 3  2017  21.5   77318     3
## 4  2017  25.9   21191     3
## 5  2016  19.2   75364     4
## 6  2017  19.2   71433     3
```

Model 1: Use Age and Miles as predictors

a.) Fit the model with two predictors (age and miles) for price as the response variable and provide the output (both the summary and the anova for the model).

Answer: Fitted model: $\widehat{Price} = 27.96 - .7425Age - .00007Mileage$ where price is in thousands of dollars.

```
mod1<- lm(Price~ Age + Mileage, data=OutbackData)
summary(mod1)
```

```
##
## Call:
## lm(formula = Price ~ Age + Mileage, data = OutbackData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8974 -1.2953 -0.2542  1.2070  3.8725
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.796e+01  4.353e-01  64.240  < 2e-16 ***
## Age         -7.425e-01  1.096e-01  -6.776 8.10e-09 ***
## Mileage     -7.798e-05  7.734e-06 -10.084 3.38e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.77 on 56 degrees of freedom
## Multiple R-squared:  0.9148, Adjusted R-squared:  0.9117
## F-statistic: 300.5 on 2 and 56 DF,  p-value: < 2.2e-16
```

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 1564.53 1564.53  499.34 < 2.2e-16 ***
## Mileage    1  318.59  318.59  101.68 3.376e-14 ***
## Residuals 56  175.46    3.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.) Find the residual for the first car in your sample. Show the actual computation for this part, based on your prediction equation and the data for that car.

Answer:

Residual = Actual - Fitted

Actual = 20.45 Fitted: $27.96 - .7425(3) - .00007(37568) = 23.10274$

Residual = $20.45 - 23.10274 = -2.652$

c.) Assess the importance of each of the predictors in the model: be sure to indicate the specific value(s) from the output you are using to make the assessments. Include hypotheses and conclusions in context.

Answer:

```
mod_PA<- lm(Price~Age, data=OutbackData)
summary(mod_PA)
```

```
##
## Call:
## lm(formula = Price ~ Age, data = OutbackData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.523 -2.000 -0.178  2.071  5.595
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.8816     0.7017   38.31   <2e-16 ***
## Age          -1.5847     0.1180  -13.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.944 on 57 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7558
## F-statistic: 180.5 on 1 and 57 DF,  p-value: < 2.2e-16
```

```
cor.test (Price~Age, data=OutbackData)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Price and Age
## t = -13.435, df = 57, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9220241 -0.7926589
## sample estimates:
##       cor
## -0.8717818
```

```
mod_PM <- lm(Price~ Mileage, data=OutbackData)
summary(mod_PM)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = OutbackData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2883 -1.1038 -0.1912  1.0618  7.1228
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.696e+01  5.478e-01   49.23   <2e-16 ***
## Mileage     -1.179e-04  6.693e-06  -17.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.367 on 57 degrees of freedom
## Multiple R-squared:  0.8449, Adjusted R-squared:  0.8422
## F-statistic: 310.5 on 1 and 57 DF,  p-value: < 2.2e-16
```

```
cor.test(Price~Mileage, data=OutbackData)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Price and Mileage
## t = -17.62, df = 57, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9513286 -0.8672222
## sample estimates:
##        cor
## -0.9191741
```

```
cor.test(Age~Mileage, data=OutbackData)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Age and Mileage
## t = 8.892, df = 57, p-value = 2.319e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6289962 0.8520469
## sample estimates:
##       cor
## 0.7622924
```

Hypotheses: $H_0 : \beta_i = 0$ and $H_1 : \beta_i \neq 0$ Both of the predictors in the model are important. To make this assessment we use the p-value from the individual t-test for the coefficents from the first summary. We also test a model with just one of the predictor variables to see if it better explains the data then the combined model. We can see that a model with $Price$ and Age gives an $R^2 = .76$ and a model with $Price$ and $Mileage$ gives an $R^2 = .85$ and the model with both predictor variables gives us an $R^2 = .914$ This means that both of our variables are imporatant and the combined model is better then two individual models.

d.) Assess the overall effectivness of this model (with a formal test). Again, be sure to include hypotheses and the specific value(s) you are using from the output to reach a conclusion.
Answer: Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ and H_a : at least one $\beta_i \neq 0$ We can see from the ANOVA output above, that the p-values are significant and the F-statistic is = 300.5. This gives stong evidence to reject the null hypothesis and conclude that at least one of the predictors, Age or $Mileage$ is effective for explaining variability in $Price$ of cars. With the individual t-tests above we see that both predictors are effective.

e.) Compute and interpret the variance inflation factor (VIF) for your predictors.
Answer:
Our text suggests we should be especially concerned when the variance inflation factor is greater then 5, i.e., when 80% of the variability in one of the explanatory variables is explained by the other variables in the model. Neither of our predictors has a VIF that is greater then 5, so there is concern over the Age and $Mileage$ variable. Both of our variables have VIF=2.387 meaning that both variables are not strongly related to each other.

```
vif(mod1)
```

```
##      Age  Mileage
## 2.387146 2.387146
```

Model 2: Polynomial Models

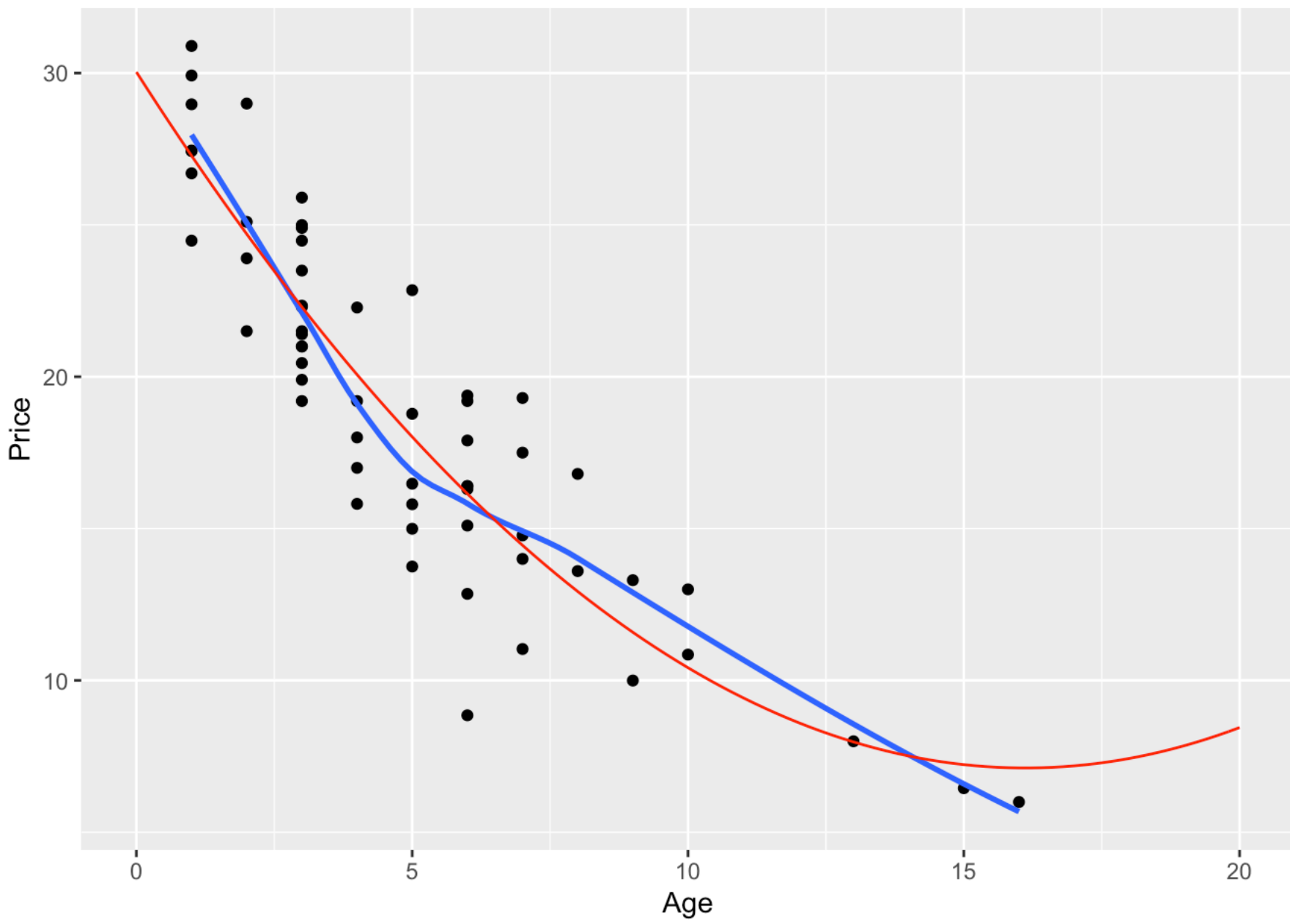
a.) Fit a quadratic model (polynomial in X and X^2) using *age* to predict price. Give the prediction equation and show a scatterplot of data with the quadratic fit drawn on it. (Hint: Use the code provided from this section in class to see how to draw it.)
Answer: The prediction equation is as follows: $\widehat{Price} = 30.034 - 2.844(Age) + .088(Age)^2$

```
mod2 <- lm(Price ~ Age + I(Age^2), data = OutbackData)
summary(mod2)
```

```
##
## Call:
## lm(formula = Price ~ Age + I(Age^2), data = OutbackData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2964 -1.5664  0.0448  1.9687  4.8505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.03417    0.97351  30.852  < 2e-16 ***
## Age         -2.84402    0.31799  -8.944 2.22e-12 ***
## I(Age^2)     0.08823    0.02106   4.190 1.00e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.592 on 56 degrees of freedom
## Multiple R-squared:  0.8173, Adjusted R-squared:  0.8108
## F-statistic: 125.2 on 2 and 56 DF,  p-value: < 2.2e-16
```

```
my_fun <- makeFun(mod2)
gf_point(Price ~ Age, data=OutbackData) %>%
        gf_smooth() %>%
        gf_fun(my_fun(Age) ~ Age, color = "red") %>%
  gf_refine(xlim(0,20))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



b.) You are looking at a seven-year-old car of your model and want to find an interval that is likely to contain its price using your quadratic model. Include the R code and interpret your interval in context in a sentence.

Answer: The interval from R code is (9.1582,19.7406) which means in context: We predict and are 95% confident that a indvidual seven-year old Subaru Outback that we are looking at would most likely cost between $9,158.20 and $19,740.06.

```
predict(mod2, data.frame(Age = 7),
        se.fit = T, interval = "prediction")
```

```
## $fit
##        fit      lwr      upr
## 1 14.44946 9.158223 19.74069
##
## $se.fit
## [1] 0.5097529
##
## $df
## [1] 56
##
## $residual.scale
## [1] 2.591681
```

c.) Does that quadratic model allow for some *age* where a car has a zero or negative predicted price? Justify your answer using a calculation or graph.
Answer: The model does not allow for an Age where the car has a zero or negative predicted price. This is becuase the model is a quadratic, meaning it has a U shape, therefore after ~16 Age in our model the graph starts to rise.(Also can be plugged into the prediction equation and we can see there are no real solutions when price is 0) This is not realistic to the price of a car, after a certain age (which we found last time ~16.5) the car is worth nothing. This can be seen by extending the x-axis on our graph above.

d.) What happens in the quadratic model for cars that are very old? Can you think of a plausible “real world” explanation for this, or is it a flaw in the quadratic model?
Answer: The price starts to rise for cars that are very old. A explanation for a small portion of these cars is that they would be worth more in parts then they are running, so maybe the price would go up in that sense. Or possibly for collectable cars, the older they get the value goes up, but only when they are in good conditon. But generally this is a flaw in our quadratic model, it is only a plausible model up to a certain age, which for us is less then 16 years old it seems.

e.) Would the fit improve significantly if you also included a cubic term? Justify your answer.
Answer: It can be seen from the nested-F test that the F-statistic is 3.249 and the correspnding p-value = .0769 which is >.05. This leads us to conclude that the quadtratic model is preferable over the cubic model. Also note that in the summary statistics of the cubic model the cubed term also has a insignificant p-value(=.0769) and the R^2 value and adjusted R^2 both decreased in the cubic model compared to the quadratic. Therefore, the fit does not improve significantly if the cubic term in included.

```
mod2.1<- lm(Price~ Age + I(Age^2)+ I(Age^3), data=OutbackData)
summary(mod2.1)
```

```
##
## Call:
## lm(formula = Price ~ Age + I(Age^2) + I(Age^3), data = OutbackData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1290 -1.4430 -0.3534  1.6900  5.2562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.068916   1.478278  21.693  < 2e-16 ***
## Age         -4.268958   0.849764  -5.024 5.69e-06 ***
## I(Age^2)     0.325961   0.133488   2.442   0.0179 *
## I(Age^3)    -0.010235   0.005678  -1.803   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.541 on 55 degrees of freedom
## Multiple R-squared:  0.8275, Adjusted R-squared:  0.8181
## F-statistic: 87.93 on 3 and 55 DF,  p-value: < 2.2e-16
```

```
anova(mod2, mod2.1)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Age + I(Age^2)
## Model 2: Price ~ Age + I(Age^2) + I(Age^3)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     56 376.14
## 2     55 355.16  1    20.982 3.2493 0.07693 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3: Complete Second-Order Model

a.) Write down the complete second order model for predicting a used car **price** based on **age** and **miles**. Note: This is the model before estimating the fit so use general coefficients β_i.

Answer: Second order model: $\widehat{Price} = \beta_0 + \beta_1(Age) + \beta_2(Age)^2 + \beta_3(Mileage) + \beta_4(Mileage)^2 + \beta_5(Age * Mileage) + \epsilon$

b.) Use R to estimate the coefficients in the model. Include summary and anova output and write down the prediction equation with proper statistical notation.

Answer: The prediction equation is:

$\widehat{Price} = 29.88 + -1.249(Age) + .018(Age)^2 + .000103(Mileage) + (1.095 * 10^{-10})(Mileage)^2 + (2.120 * 10^{-06})(Age * Mileage)$

The anova output and summary are both showing that the $Mileage$ and interaction of Age and $Mileage$ terms are not significant to the model and we may want to consider taking them out of our model.

```
mod3<- lm(Price ~ Age+I(Age^2)+Mileage + I(Mileage^2)+ I(Age*Mileage), data=OutbackData)
summary(mod3)
```

```
##
## Call:
## lm(formula = Price ~ Age + I(Age^2) + Mileage + I(Mileage^2) +
##     I(Age * Mileage), data = OutbackData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5453 -1.0750  0.0087  1.2107  2.8334
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.988e+01  5.818e-01  51.363  < 2e-16 ***
## Age              -1.249e+00  3.514e-01  -3.555 0.000805 ***
## I(Age^2)          1.804e-02  5.149e-02   0.350 0.727480
## Mileage          -1.035e-04  2.460e-05  -4.205 0.000101 ***
## I(Mileage^2)      1.095e-10  1.029e-10   1.064 0.292311
## I(Age * Mileage)  2.120e-06  4.910e-06   0.432 0.667665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.544 on 53 degrees of freedom
## Multiple R-squared:  0.9386, Adjusted R-squared:  0.9328
## F-statistic: 162.1 on 5 and 53 DF,  p-value: < 2.2e-16
```

```
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: Price
##                  Df  Sum Sq Mean Sq  F value    Pr(>F)
## Age               1 1564.53 1564.53 656.2835 < 2.2e-16 ***
## I(Age^2)          1  117.91  117.91  49.4612 3.993e-09 ***
## Mileage           1  243.72  243.72 102.2368 5.707e-14 ***
## I(Mileage^2)      1    5.62    5.62   2.3592    0.1305
## I(Age * Mileage)  1    0.44    0.44   0.1864    0.6677
## Residuals        53  126.35    2.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c.) Show the details of a nested F-test for the importance of the second-order terms (quadratic and interaction) that involve **miles** in this model. Include the output needed to get the information to complete the test. Show all the usual details (hypotheses, test statistic, p-value, and an informative conclusion in context). In particular, show how the nested F test statistic is computed by putting values into a formula.

Answer:

Hypotheses: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_5 = 0$ and H_1 : at least one $\beta \neq 0$.

The test statistic is $F = \frac{SSE_R - SSE_F / DF_F}{SSE_F / [n-(k+p+1)]}$ Our F test statistic: $F = \frac{431.04/3}{126.35/53} = 60.271$ From the nested F-test below we can see that out F = 60.271 and the p-value is less then .001. This would lead us to conclude that the full model is better, and at least one of the predictors that we took out for the reduced model is useful. From the summary we can see that the R^2 went down from (.91 to .72) for the reduced model. The F-statistic also went down from (162 to 75). Therefore we conclude that the full model is preferable. In context this means that all the predictors of Age, Age^2, Mileage, Mileage^2 and Age*Mileage are better at predicting the price of a Subaru Outback compared to a reduced version.

```
mod_full <- lm(Price ~ Age + I(Age^2) + Mileage + I(Mileage^2) + I(Age*Mileage), data=OutbackData)
vif(mod_full)
```

```
##              Age         I(Age^2)          Mileage     I(Mileage^2)
##         32.27810        157.98589         31.74692          18.19135
## I(Age * Mileage)
##        182.60119
```

```
mod_reduced <- lm(Price ~ I(Mileage^2) + I(Age*Mileage), data=OutbackData)
summary(mod_reduced)
```

```
##
## Call:
## lm(formula = Price ~ I(Mileage^2) + I(Age * Mileage), data = OutbackData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5020 -2.3156 -0.9204  1.7828  7.6801
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.322e+01  5.369e-01  43.259  < 2e-16 ***
## I(Mileage^2)     -2.254e-10  9.418e-11  -2.394   0.0201 *
## I(Age * Mileage) -6.050e-06  1.418e-06  -4.266 7.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.155 on 56 degrees of freedom
## Multiple R-squared:  0.7292, Adjusted R-squared:  0.7196
## F-statistic: 75.41 on 2 and 56 DF,  p-value: < 2.2e-16
```

```
anova(mod_reduced, mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ I(Mileage^2) + I(Age * Mileage)
## Model 2: Price ~ Age + I(Age^2) + Mileage + I(Mileage^2) + I(Age * Mileage)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     56 557.39
## 2     53 126.35  3    431.04 60.271 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wrap-Up

From the various models that we have considered throughout this project I would reccommend using the full-second order model for predicting price of the car. My justification for this model is that the $R^2 = 0.9386$ and adjusted $R^2 = 0.9328$ is the highest. The F-value=60.271 is high when comparing with a reduced version and the F-Statistic=75.41 is smallest for our full model, finally the full model has the smallest standard error = 3.155. Therefore using a full-second order model with Age and $Mileage$ for predicting the $Price$ of a car is the best choice, all of the predictor variables have a significant effect in the price of a Subarbu Outback.