**STOR 455 Homework #4**

**40 points - Due on 10/13 at 5:00pm**

**Situation:** Suppose that (again) you are interested in purchasing a used car. How much should you expect to pay? Obviously the price will depend on the type of car you get (the model) and how much it’s been used. For this assignment you will investigate how the price might depend on the age and mileage, as well as the state where the car is purchased.

**Data Source:** To get a sample of cars, begin with the UsedCars CSV file. The data was acquired by scraping TrueCar.com for used car listings on 9/24/2017 and contains more than 1.2 million used cars. For this assignment you should choose the same car *Model* and *State* that you initially chose for homework #2. You should again add a variable called *Age* which is 2017-year (since the data was scraped in 2017).

**Directions:** The code below can again be used to select data from a particular *Model* and *State* of your choice. The R chunk below begins with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you run this chunk, you should revert it to {r}.

library(readr)

library(car)

ModelOfMyChoice = "Civic"

StateOfMyChoice = "NY"

UsedCars <- read\_csv("UsedCars.csv")

MyCars = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice)

MyCars$Age = 2017 - MyCars$Year

**MODEL #4: Use Age and Miles as predictors for Price**

1. Construct a model using two predictors (age and miles) with *Price* as the response variable and provide the summary output.

**0.5 points** model  
**0.5 points** summary output

modq1 = lm(Price~Age+Mileage, data=MyCars)

summary(modq1)

##

## Call:

## lm(formula = Price ~ Age + Mileage, data = MyCars)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4694.7 -1443.3 -316.7 1099.8 7168.3

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 1.817e+04 1.660e+02 109.499 < 2e-16 \*\*\*

## Age -1.036e+03 6.319e+01 -16.396 < 2e-16 \*\*\*

## Mileage -2.543e-02 4.642e-03 -5.478 7.23e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1901 on 445 degrees of freedom

## Multiple R-squared: 0.7325, Adjusted R-squared: 0.7313

## F-statistic: 609.4 on 2 and 445 DF, p-value: < 2.2e-16

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

**2 point** (1 pt each) They should use the summary of the model to comment on the p-values for the individual slope tests. They do not need to specifically list the hypotheses being tested, just note if the p-values are small. If they instead do tests for correlation, I’ll allow that for full credit as well if they again comment on the p-values without the need to cite hypotheses. For my model both p-values for the *Age* and *Mileage* predictors are well below 0.05, hence useful in the model.

1. Assess the overall effectiveness of this model (with a formal test). Again, be sure to include hypotheses and the specific value(s) you are using from the summary output to reach a conclusion.

**2 points** I expect them to perform a hypothesis test with Null: βi = 0 for all i, Alternative βi ≠ 0 for some i. and draw a conclusion from the p-value of anova455() or the similar output in the summary table. They can write out these conclusion in words, such as the null hypothesis is that all coefficients are zero, the alternative is that at least one is nonzero. If p-values or significance are mentioned in the hypotheses, deduct 0.5 points. For my model the p-value is small (2.2e-16), so I have evidence to support the alternative, that at least one of the coefficients is nonzero.

**Note:** They should not draw this conclusion from the p-values in the anova() output. By default this uses a sequential sums method, which performs a series of nested F tests. You can take off 1 point for using anova() instead of anova455() or the summary() (Summary is also full credit).

**Note:** Throughout the assignment, anova455() may show different outputs in students’ notebooks vs the knitted html. In the notebook, the values will likely not go below 2.2e-16. When knit, this does not seem to be the case, If you see this come up, make sure **not** to take off credit when the students cite 2.2e-16 as the p-value, even the the output might show a lower number.

source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")

anova455(modq1)

|  |
| --- |
|  |

|  | **Df**  **<dbl>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **P(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Model | 2 | 4402719393 | 2201359696 | 609.3773 | 0 |
| Error | 445 | 1607551043 | 3612474 | NA | NA |
| Total | 447 | 6010270436 | NA | NA | NA |

3 rows

summary(modq1)

##

## Call:

## lm(formula = Price ~ Age + Mileage, data = MyCars)

##

## Residuals:

## Min 1Q Median 3Q Max

## -4694.7 -1443.3 -316.7 1099.8 7168.3

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 1.817e+04 1.660e+02 109.499 < 2e-16 \*\*\*

## Age -1.036e+03 6.319e+01 -16.396 < 2e-16 \*\*\*

## Mileage -2.543e-02 4.642e-03 -5.478 7.23e-08 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1901 on 445 degrees of freedom

## Multiple R-squared: 0.7325, Adjusted R-squared: 0.7313

## F-statistic: 609.4 on 2 and 445 DF, p-value: < 2.2e-16

1. Compute and interpret the variance inflation factor (VIF) for your predictors.

**1 point** compute VIF - most will use the vif() function from the car package. This package has issues intalling on some macs, so I also made available a script of this function, which they may use as well. It’s possible that they calculate it from the R2 of the model using Predictor1~Predictor2, then vif = 1/(1-R2). This is fine as well.

**1 point** discuss VIF - They should in some may say that there is little, or substantial multicollinearity based on the VIF value. We haven’t concrete cutoffs for this, but over 5 may be substantial, with lower as little multicollinearity. Here my VIF is fairly small, so there is little concern about multicollinearity.

vif(modq1)

## Age Mileage

## 2.756156 2.756156

1. Suppose that you are interested in purchasing a car of this model that is four years old (in 2017) with 31K miles. Determine each of the following: a 90% confidence interval for the mean price at this age and mileage, and a 90% prediction interval for the price of an individual car at this age and mileage. Write sentences that carefully interpret each of the intervals (in terms of car prices)

**1 points** new dataframe for 4 year old car with 31000 miles  
**0.5 points** confidence interval  
**0.5 points** prediction interval  
**1 point** (0.5 each) They should clearly distinguish that the confidence interval is predicting the mean price of four year old cars with 31K miles in their model, while the prediction interval is predicting the price of a specific car of their model that is four years old with 31K miles. The textbook also describes the prediction interval interval as predicting the interval where most cars of this age/model would be contained. This is fine as well.

oneCar2 = data.frame(Age = 4, Mileage=31000)

predict.lm(modq1, oneCar2, interval = "confidence", level=.9)

## fit lwr upr

## 1 13242.38 13064.6 13420.16

predict.lm(modq1, oneCar2, interval = "prediction", level=.9)

## fit lwr upr

## 1 13242.38 10104.52 16380.23

**MODEL #5: Now Include a Categorical predictor**

For this section you will combine both datasets used in Homework #2, as well as a third dataset. Each dataset from Homework #2 included cars from your specific *Model*, but from two different states. You should use the same code that you used in homework #2 to construct this second dataframe with cars from North Carolina, and a third dataframe with cars of your model from a third state of your choice. Then manipulate the code below to combine the three dataframes into one dataframe. Make sure to add the *Age* variable again to your dataframes for the additional states before binding them together. The R chunk below begins with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you run this chunk, you should revert it to {r}.

MyCars2 = subset(UsedCars, Model==ModelOfMyChoice & State=="NC")

MyCars3 = subset(UsedCars, Model==ModelOfMyChoice & State=="CA")

MyCars2$Age = 2017 - MyCars2$Year

MyCars3$Age = 2017 - MyCars3$Year

State1 = MyCars

State2 = MyCars2

State3 = MyCars3

# rbind combines the rows in one dataframe, assuming that the columns are the same.

CombinedStates = rbind(State1, State2, State3)

1. Fit a multiple regression model using *Age*, *Mileage*, and *State* to predict the *Price* of the car.

**1 pt** - code for model. They may factor() *State*, but this is redundant.

modq6 = lm(Price~Age+Mileage+State, data=CombinedStates)

1. Perform a hypothesis test to determine the importance of *State* terms in the model constructed in question 6. List your hypotheses, p-value, and conclusion.

**2 pt** - Construct a reduced model and use anova().  
**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (two) State terms. Take off 0.5 points if they state that there is only one coefficient being tested.  
**0.5** pts - conclusion

H0: β3 = β4A: β3 ≠ 0 or β4 ≠ 0;

Reject the null. There is statistically significant evidence (2.127e-12) to suggest that at least one coefficient of a State variable is nonzero.

modq7 = lm(Price~Age+Mileage, data=CombinedStates)

anova(modq7, modq6)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 11752614042 | NA | NA | NA | NA |
| 2 | 2487 | 11501323368 | 2 | 251290674 | 27.16904 | 2.126612e-12 |

2 rows

1. Fit a multiple regression model using *Age*, *Mileage*, *State*, and the interactions between *Age* and *State*, and *Mileage* and *State* to predict the *Price* of the car.

**2 pt** - code for model. They may factor() *State*, but this is redundant . Note that if they only include the interaction terms, the lm function will ‘fill in the blanks’ for them and create a model using the individual terms as well. So Price ~ both interaction terms would produce the correct model as well.

modq8 = lm(Price~Age+Mileage+State + Age\*State + Mileage \* State, data=CombinedStates)

# Also correct

modq8.1 = lm(Price~Age\*State + Mileage \* State, data=CombinedStates)

1. Perform a hypothesis test to determine the importance of *State* terms in the model constructed in question 8. List your hypotheses, p-value, and conclusion.

**1 pt** - hypotheses. Take off 0.5 points if they state that there is only two coefficients being tested.  
**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (six) Model terms.  
**0.5 pts** - conclusion

H0: βi = 0; for all i, i=(3,4,5,6,7)

HA: βi ≠ 0; for at least one i, i=(3,4,5,7)

The 3rd through 7th terms of the model contain a State term.

Reject the null. There is statistically significant evidence (9.002e-15) to suggest that at least one of the coefficients for a term with State in the linear model is nonzero.

modq9 = lm(Price~Age+Mileage, data=CombinedStates)

anova(modq9, modq8)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 11752614042 | NA | NA | NA | NA |
| 2 | 2483 | 11389191105 | 6 | 363422937 | 13.2052 | 9.001764e-15 |

2 rows

**MODEL #6: Polynomial models**

One of the drawbacks of the linear model in homework #2 was the “free car” phenomenon where the predicted price is eventually negative as the line decreases for older cars. Let’s see if adding one or more polynomial terms might help with this. For this section you should use the dataset with cars from three states that you used for model 5.

1. Fit a quadratic model using *Age* to predict *Price* and examine the residuals. Construct a scatterplot of the data with the quadratic fit included. You do not need to specifically cite all conditions for the linear model, but should discuss any issues that you see in the conditions.

**1.5 pt** - code for quadratic model (may use poly function or create new squared vairable in the dataframe)  
**2 pt** - plot with quadratic curve  
**1 pt** - conditions for model. They do not need to comment on all of the conditions for the linear model. You can give the full point for some discussion of the residuals in terms of the 3 conditions. This could be as simple as saying that for my data most of the conditions look good, with the possible the qqnorm plot showing a slight deviation from the qqline at the right tail, which would impact the normality of the residuals.

modq10 = lm(Price~Age+I(Age^2), data=CombinedStates)

# alternative method using the poly() function

# Must have Raw=TRUE or the two methods will not be the same

modq10poly = lm(Price~poly(Age, degree=2, raw=TRUE), data=CombinedStates)

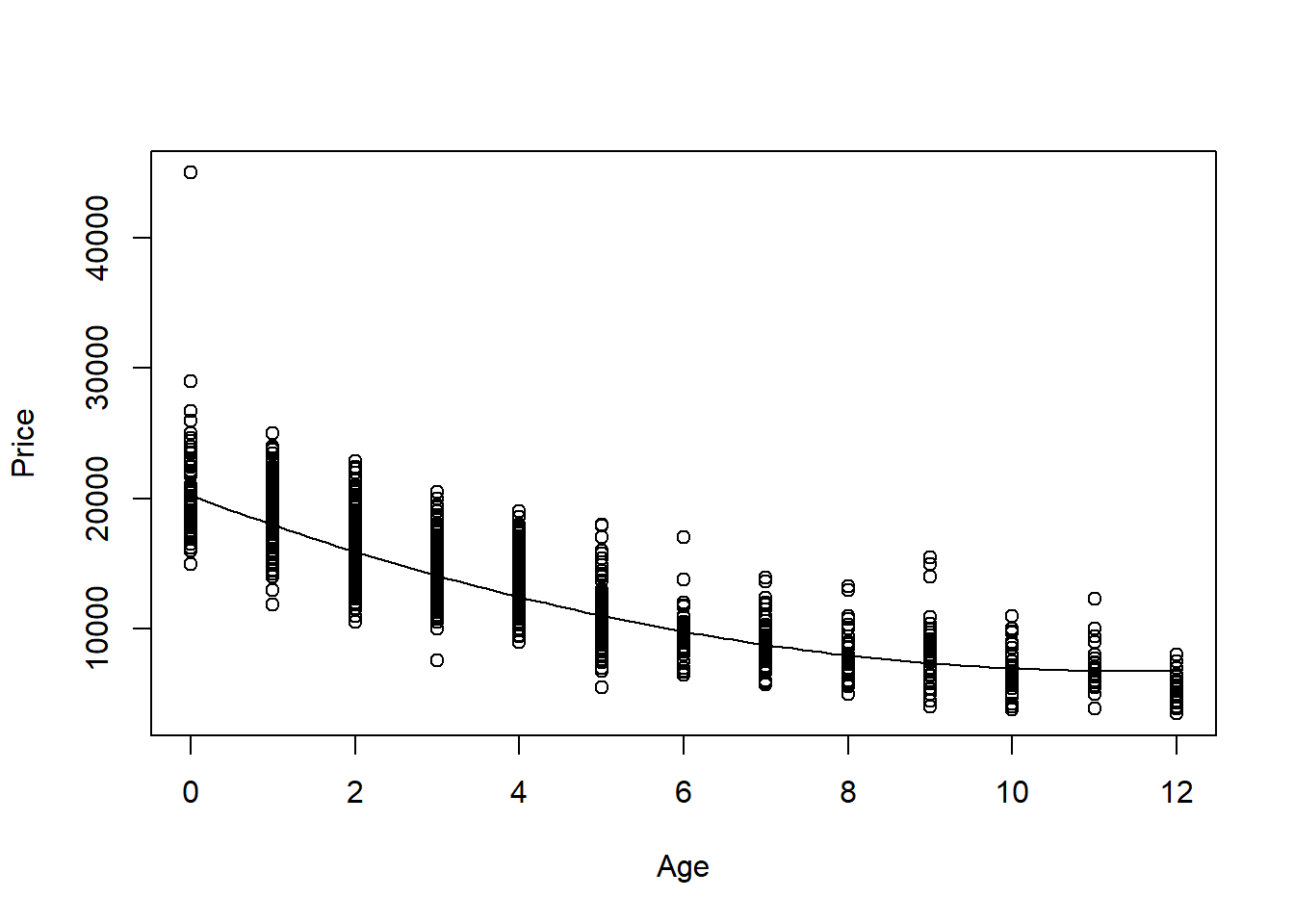
plot(Price~Age, data=CombinedStates)

a = summary(modq10)$coef[3]

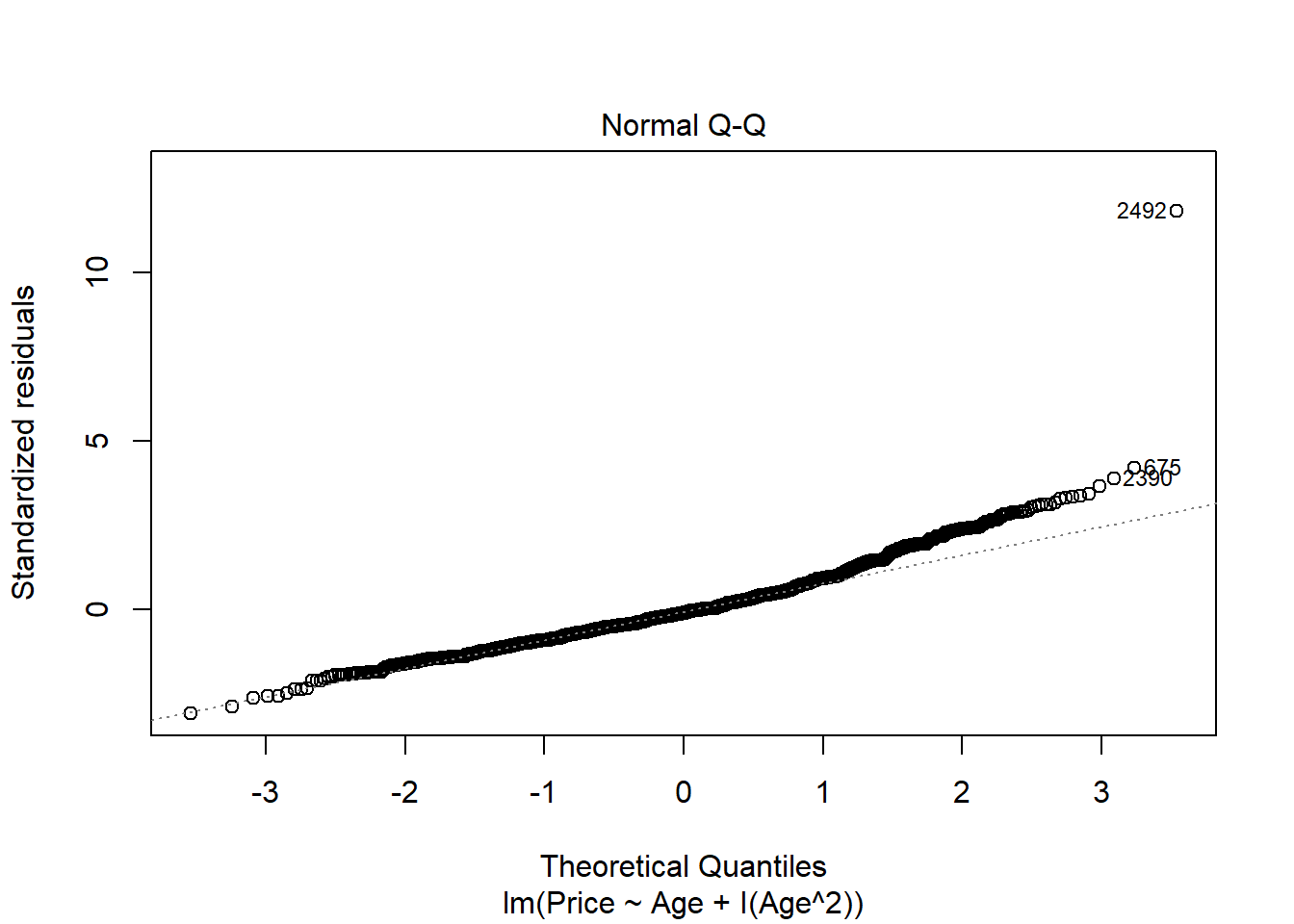
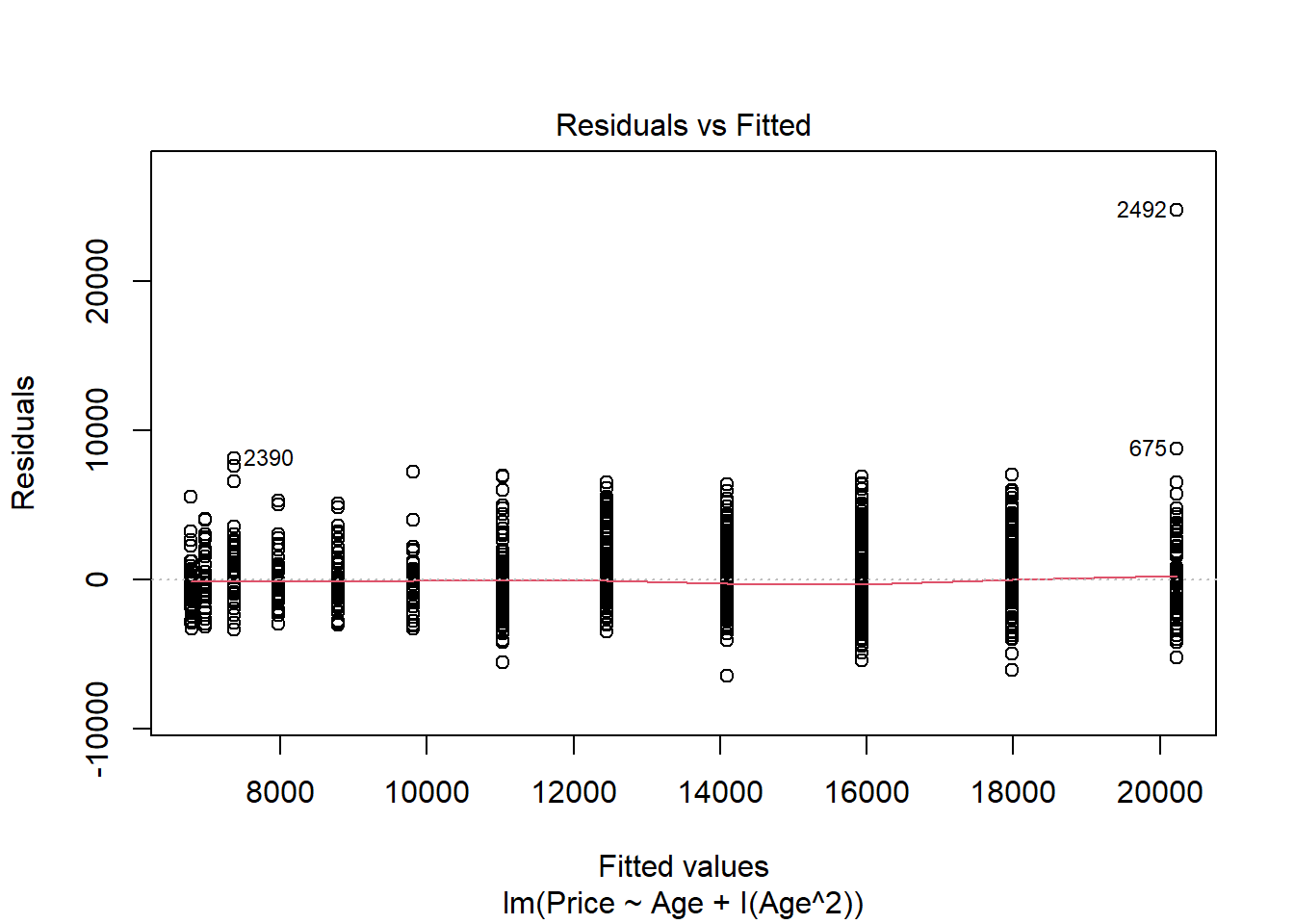
b = summary(modq10)$coef[2]

c = summary(modq10)$coef[1]

curve(a\*x^2 + b\*x + c, add=TRUE)



plot(modq10, c(1,2))



1. Perform a hypothesis test to determine if this model is significant. List your hypotheses, p-value, and conclusion.

**1 pt** - hypotheses  
**0.5 pts** - anova455 or anova test from summary to get the p-value if they built their model without using poly(). If they use anova() on a non poly() model, they are not doing the correct test. If they use anova() on a poly() model, the result is correct.  
**0.5 pts** - conclusion

H0: βi = 0; for all i

HA: βi ≠ 0; for at least one i

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one coefficient in the model is nonzero.

anova455(modq10)

|  |
| --- |
|  |

|  | **Df**  **<dbl>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **P(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Model | 2 | 29206450923 | 14603225462 | 3330.633 | 0 |
| Error | 2489 | 10913068062 | 4384519 | NA | NA |
| Total | 2491 | 40119518985 | NA | NA | NA |

3 rows

# or

summary(modq10)

##

## Call:

## lm(formula = Price ~ Age + I(Age^2), data = CombinedStates)

##

## Residuals:

## Min 1Q Median 3Q Max

## -6491.9 -1347.2 -238.2 1022.5 24760.8

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 20227.163 114.134 177.22 <2e-16 \*\*\*

## Age -2353.440 54.602 -43.10 <2e-16 \*\*\*

## I(Age^2) 102.785 4.843 21.23 <2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 2094 on 2489 degrees of freedom

## Multiple R-squared: 0.728, Adjusted R-squared: 0.7278

## F-statistic: 3331 on 2 and 2489 DF, p-value: < 2.2e-16

# or

anova(modq10poly)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| poly(Age, degree = 2, raw = TRUE) | 2 | 29206450923 | 14603225462 | 3330.633 | 0 |
| Residuals | 2489 | 10913068062 | 4384519 | NA | NA |

2 rows

# incorrect

anova(modq10)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Age | 1 | 27231216805 | 27231216805 | 6210.7648 | 0.000000e+00 |
| I(Age^2) | 1 | 1975234118 | 1975234118 | 450.5019 | 4.971935e-92 |
| Residuals | 2489 | 10913068062 | 4384519 | NA | NA |

3 rows

1. You are looking at a 4-year-old car of your model and want to find an interval that is likely to contain its *Price* using your quadratic model. Construct an interval to predict the value of this car, and include an interpretive sentence in context.

**1 pts** - new dataframe wth age=4 car  
**0.5 pts** - prediction interval at any confidence level (not confidence interval!)  
**0.5 pts** - conclusion specific to the prediction of this one particular car’s price

ThisCar = data.frame(Age=4)

predict.lm(modq10, ThisCar, interval="prediction")

## fit lwr upr

## 1 12457.97 8350.305 16565.63

1. Does the quadratic model allow for some *Age* where a car has a zero or negative predicted price? Justify your answer using a calculation or graph.

**2 pt** - yes/no with some justification. Some students had concave down parabalas, which they could note would clearly go below zero. Other have concave up, and may need to in some way find the roots of the equation, or plot the curve in some way that it is clear if it crosses below the horizontal axis.

Note: We did not use the polyroot function in class, so they likely found the roots some other way. The roots are imaginary, so the Price never goes below zero. This is also seen in the plot below. My roots are imaginary, showing that the parabol never crosses zero.

# shows only imaginary roots, hence Price never equals 0

polyroot(c(c, b, a))

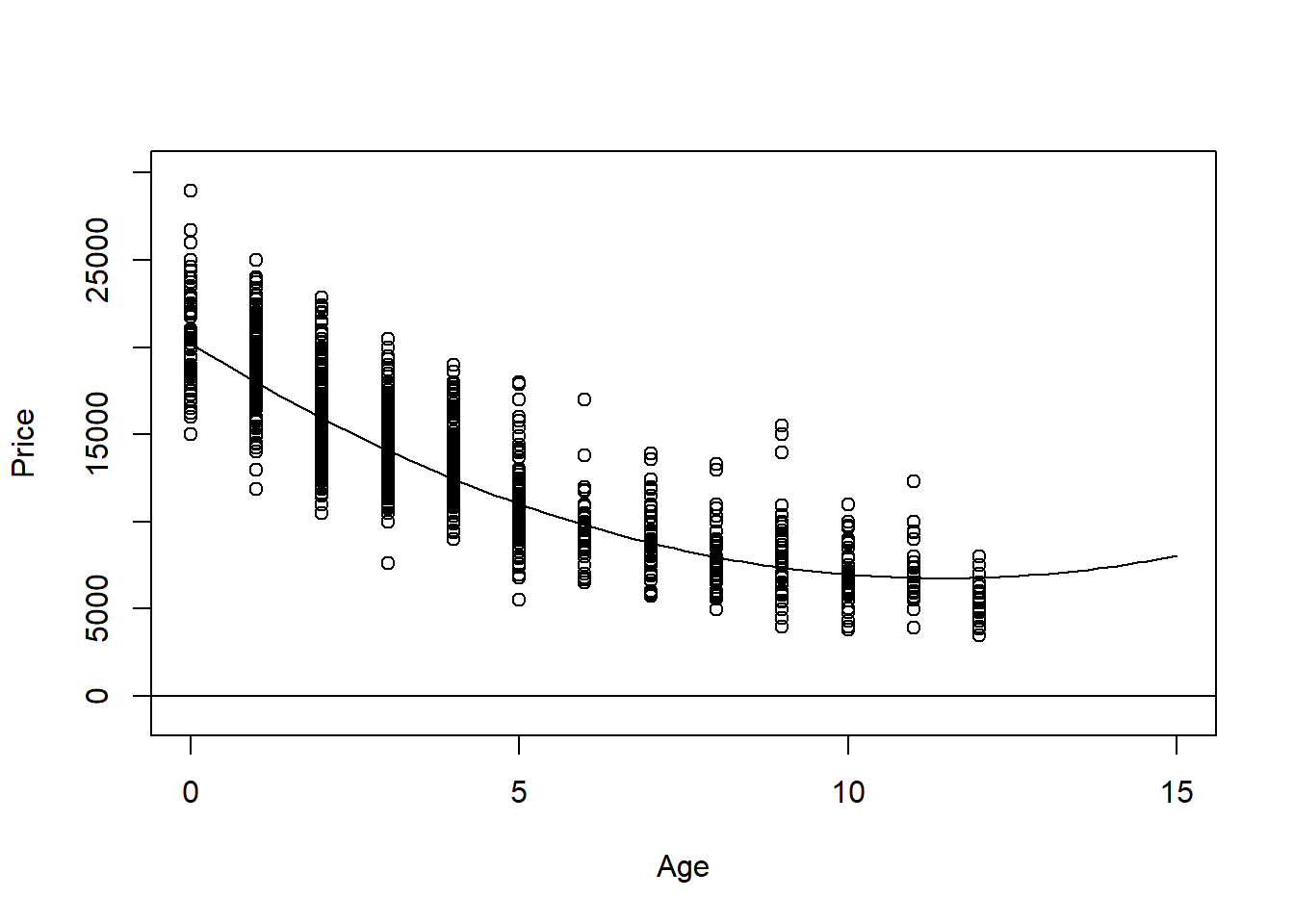
## [1] 11.44833+8.10717i 11.44833-8.10717i

# Or plot

plot(Price~Age, data=CombinedStates, xlim=c(0,15), ylim=c(-1000, 30000))

curve(a\*x^2 + b\*x + c, add=TRUE)

abline(0,0)



1. Would the fit improve significantly if you also included a cubic term? Does expanding your polynomial model to use a quartic term make significant improvements? Justify your answer.

**2 pts** - There are many ways that students could reasonably answer this. They do not need to note specific hypotheses for hypothesis tests, but they should draw their conclusion from performing a hypothesis test. This could be with a nested F test for quadratic and cubic models, quadratic and quartic models, or cubic and quartic with the anova() function. using anova(quartic model) would also perform these tests for students to interpret. If they only add one term at a time, this same p-value can be found in the summary and coefficients table as well. For my data, the p-value is small for the nested tests showing the addition of the cubic term (but not quartic). THe cubic term significantly improves this model.

modq14c = lm(Price~Age+I(Age^2)+I(Age^3), data=CombinedStates)

modq14q = lm(Price~Age+I(Age^2)+I(Age^3)+I(Age^4), data=CombinedStates)

anova(modq10, modq14c)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 10913068062 | NA | NA | NA | NA |
| 2 | 2488 | 10801715538 | 1 | 111352525 | 25.64825 | 4.396797e-07 |

2 rows

anova(modq10, modq14q)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 10913068062 | NA | NA | NA | NA |
| 2 | 2487 | 10800874898 | 2 | 112193164 | 12.91675 | 2.6258e-06 |

2 rows

anova(modq14c, modq14q)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2488 | 10801715538 | NA | NA | NA | NA |
| 2 | 2487 | 10800874898 | 1 | 840639.8 | 0.193565 | 0.6600042 |

2 rows

# OR

# Check if addition of quartic term to cubic model is significant, etc...

summary(modq14q)

##

## Call:

## lm(formula = Price ~ Age + I(Age^2) + I(Age^3) + I(Age^4), data = CombinedStates)

##

## Residuals:

## Min 1Q Median 3Q Max

## -6362.6 -1338.7 -191.7 1036.4 24322.2

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 20665.7920 175.4985 117.755 <2e-16 \*\*\*

## Age -2777.4334 203.2146 -13.667 <2e-16 \*\*\*

## I(Age^2) 191.4913 76.1030 2.516 0.0119 \*

## I(Age^3) -2.8970 10.3100 -0.281 0.7787

## I(Age^4) -0.1982 0.4505 -0.440 0.6600

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 2084 on 2487 degrees of freedom

## Multiple R-squared: 0.7308, Adjusted R-squared: 0.7303

## F-statistic: 1688 on 4 and 2487 DF, p-value: < 2.2e-16

**MODEL #7: Complete second order model**

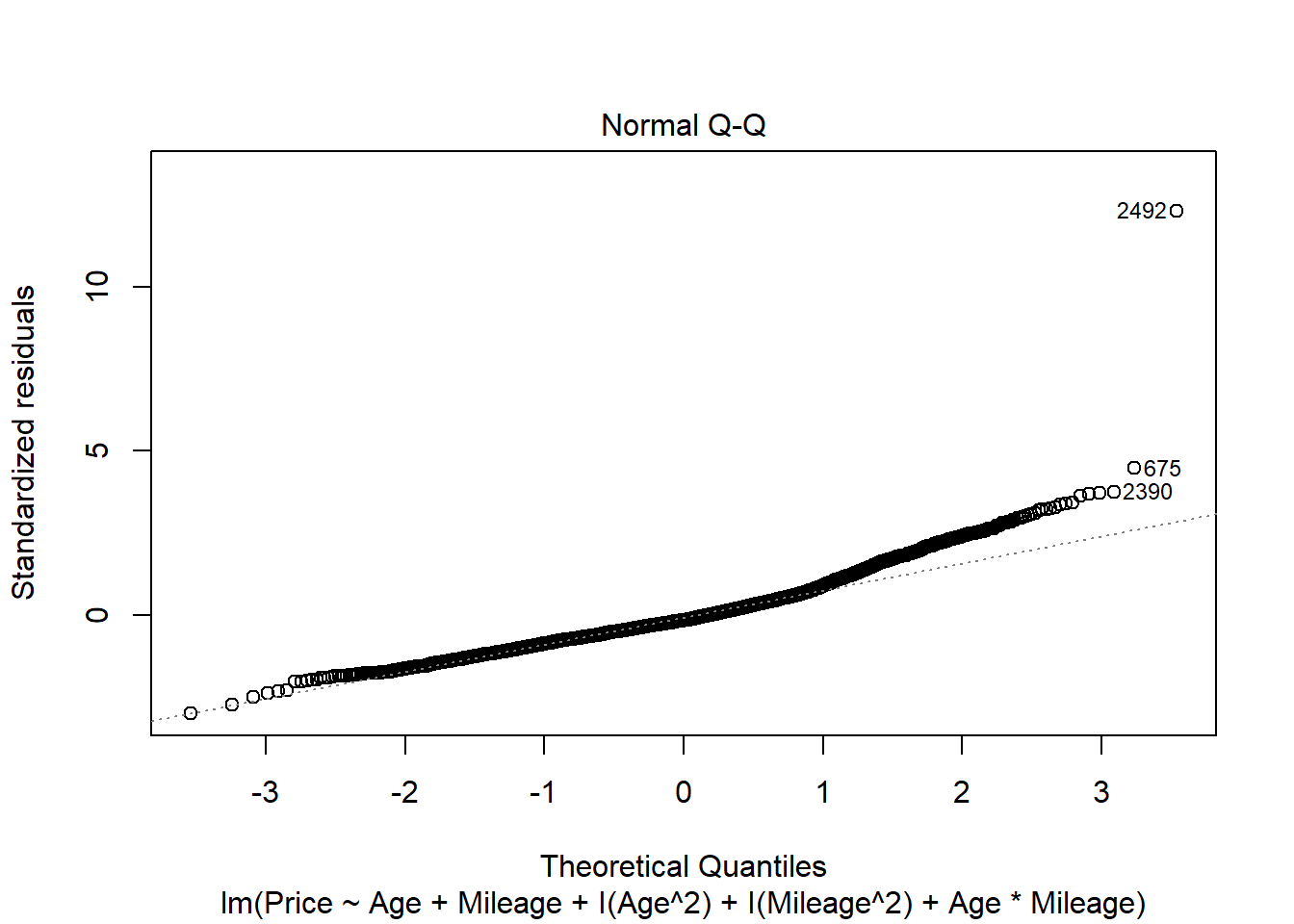
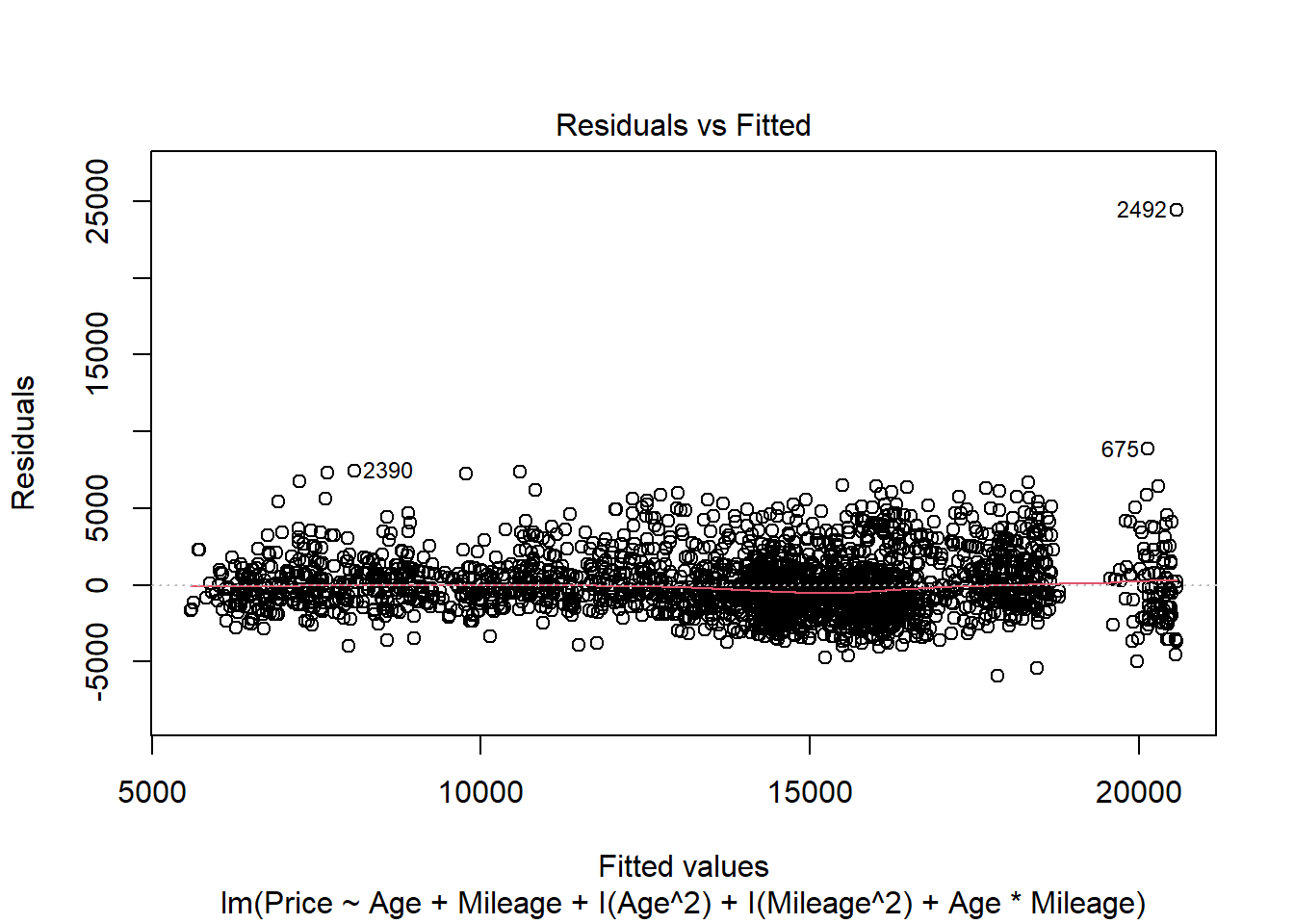
For this section you should again use the dataset with cars from three states that you used for models 5 and 6.

1. Fit a complete second order model for predicting a used car *Price* based on *Age* and *Mileage* and examine the residuals. You do not need to specifically cite all conditions for the linear model, but should discuss any issues that you see in the conditions.

**1.5 pt** - code for model. may also use polym() function although I did not do this in class.  
**1 pt** - conditions for model. They do not need to comment on all of the conditions for the linear model. You can give the full point for some discussion of the residuals in terms of the 3 conditions. As for my earlier model, the conditions seem decently met with the normality of the residuals still being a bit problematic due to the right tail deviating from the qqline.

modq15 = lm(Price~Age+Mileage+I(Age^2)+I(Mileage^2)+Age\*Mileage, data=CombinedStates)

plot(modq15, c(1,2))



# or

modq15poly = lm(Price~polym(Age, Mileage, degree=2, raw=TRUE), data=CombinedStates)

1. Perform a hypothesis test to determine if the model constructed in question 15 is significant. List your hypotheses, p-value, and conclusion.

**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all terms.  
**0.5 pts** - anova455 or anova test from summary to get the p-value if function was made without polym(). If the model is made from polym(), then the anova() function can be used.  
**0.5 pts** - conclusion

H0: βi = 0; for all i

HA: βi ≠ 0; for at least one i

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one coefficient in the model is nonzero.

anova455(modq15)

|  |
| --- |
|  |

|  | **Df**  **<dbl>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **P(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Model | 5 | 30335970693 | 6067194139 | 1541.674 | 0 |
| Error | 2486 | 9783548292 | 3935458 | NA | NA |
| Total | 2491 | 40119518985 | NA | NA | NA |

3 rows

#or

summary(modq15)

##

## Call:

## lm(formula = Price ~ Age + Mileage + I(Age^2) + I(Mileage^2) +

## Age \* Mileage, data = CombinedStates)

##

## Residuals:

## Min 1Q Median 3Q Max

## -5955.4 -1272.1 -315.6 938.5 24426.2

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.057e+04 1.150e+02 178.839 < 2e-16 \*\*\*

## Age -1.862e+03 6.284e+01 -29.629 < 2e-16 \*\*\*

## Mileage -4.556e-02 3.983e-03 -11.440 < 2e-16 \*\*\*

## I(Age^2) 8.830e+01 7.222e+00 12.225 < 2e-16 \*\*\*

## I(Mileage^2) 8.971e-08 1.719e-08 5.220 1.94e-07 \*\*\*

## Age:Mileage 1.869e-04 6.011e-04 0.311 0.756

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 1984 on 2486 degrees of freedom

## Multiple R-squared: 0.7561, Adjusted R-squared: 0.7556

## F-statistic: 1542 on 5 and 2486 DF, p-value: < 2.2e-16

#or

anova(modq15poly)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| polym(Age, Mileage, degree = 2, raw = TRUE) | 5 | 30335970693 | 6067194139 | 1541.674 | 0 |
| Residuals | 2486 | 9783548292 | 3935458 | NA | NA |

2 rows

#incorrect

anova(modq15)

|  |
| --- |
|  |

|  | **Df**  **<int>** | **Sum Sq**  **<dbl>** | **Mean Sq**  **<dbl>** | **F value**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- |
| Age | 1 | 27231216805 | 27231216805 | 6.919453e+03 | 0.000000e+00 |
| Mileage | 1 | 1135688138 | 1135688138 | 2.885784e+02 | 2.559451e-61 |
| I(Age^2) | 1 | 1838648320 | 1838648320 | 4.672006e+02 | 4.338885e-95 |
| I(Mileage^2) | 1 | 130037013 | 130037013 | 3.304241e+01 | 1.011954e-08 |
| Age:Mileage | 1 | 380417 | 380417 | 9.666399e-02 | 7.558964e-01 |
| Residuals | 2486 | 9783548292 | 3935458 | NA | NA |

6 rows

1. Perform a hypothesis test to determine the importance of just the second order terms (quadratic and interaction) in the model constructed in question 15. List your hypotheses, p-value, and conclusion.

**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (three) second order terms.  
**0.5 pts** - anova() nested test code  
**0.5 pts** - conclusion

Note: If students construct an incorrect model, but perform the test correctly on that model, they could receive points for all parts except for constructing the model.

H0: βi = 0; for all i, i=(3,4,5)

HA: βi ≠ 0; for at least one i, i=(3,4,5)

The 3rd, 4th, and 5th terms of the model are the second order terms.

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one of the second order coefficients in the model is nonzero.

modq17 = lm(Price~Age+Mileage, data=CombinedStates)

anova(modq17, modq15)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 11752614042 | NA | NA | NA | NA |
| 2 | 2486 | 9783548292 | 3 | 1969065750 | 166.7799 | 1.669608e-98 |

2 rows

1. Perform a hypothesis test to determine the importance of just the terms that involve *Mileage* in the model constructed in question 15. List your hypotheses, p-value, and conclusion.

**1 pt** - (0.5 points each) - hypotheses - could be in symbolic form as below, or in words citing the coefficients for all (three) Mileage terms.  
**0.5 pts** - anova() nested test code  
**0.5 pts** - conclusion

H0: βi = 0; for all i, i=(2,4,5)

HA: βi ≠ 0; for at least one i, i=(2,4,5)

The 2nd, 4th, and 5th terms of the model contain a Mileage term.

Reject the null. There is statistically significant evidence (p-value=2.2e-16) to suggest that at least one of the coefficients for a term with Mileage in the model is nonzero.

anova(modq10, modq15)

|  |
| --- |
|  |

|  | **Res.Df**  **<dbl>** | **RSS**  **<dbl>** | **Df**  **<dbl>** | **Sum of Sq**  **<dbl>** | **F**  **<dbl>** | **Pr(>F)**  **<dbl>** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2489 | 10913068062 | NA | NA | NA | NA |
| 2 | 2486 | 9783548292 | 3 | 1129519770 | 95.67034 | 1.342345e-58 |

2 rows