STOR 455 Homework #4

40 points - Due on 3/4 at 5:00pm

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

Data Source: To get a sample of vehicles, begin with the UsedCars CSV file. The data was acquired by scraping TrueCar.com for used vehicle listings on 9/24/2017 and contains more than 1.2 million used vehicles. For this assignment you should choose the same vehicle *Model* from North Carolina that you initially chose for homework #2.

Directions: The code below can again be used to select data from a particular *Model* of your choice from North Carolina. 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}.

```
# Delete the *** below and enter the model from homework #2
ModelOfMyChoice = "***"
StateOfMyChoice = "NC"

# Takes a subset of your model vehicle from your state
MyVehicles = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice)
```

MODEL #4: Use Year and Miles as predictors for Price

1. Construct a model using two predictors (*Year* and *Mileage*) with *Price* as the response variable and provide the summary output.
2. 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.
3. 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.
4. Compute and interpret the variance inflation factor (VIF) for your predictors.
5. Suppose that you are interested in purchasing a vehicle of this model that was four years old (in 2017) with 58K miles. Determine each of the following: a 90% confidence interval for the mean price at this Year and mileage, and a 90% prediction interval for the price of an individual vehicle at this Year and mileage. Write sentences that carefully interpret each of the intervals (in terms of vehicle prices)

MODEL #5: Now Include a Categorical predictor For this section you will combine both datasets used in Homework #2, as well as two new datasets. Each dataset from Homework #2 included vehicles 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 vehicles from the state of your choice, and a third and fourth dataframe with vehicles of your model from a third and fourth state (Choose either Arizona, Florida, or Ohio for the two additional states). Then manipulate the code below to combine the four dataframes into one dataframe. 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}.

```
State1 = MyVehicles
State2 = *** #fill in with the dataframe of cars of your model from state 2
State3 = *** #fill in with the dataframe of cars of your model from state 3
State4 = *** #fill in with the dataframe of cars of your model from state 4

# rbind combines the rows in one dataframe, assuming that the columns are the same.
CombinedStates = rbind(State1, State2, State3, State4)
```

6. Fit a multiple regression model using *Year*, *Mileage*, and *State* to predict the *Price* of the vehicle.
7. Perform a hypothesis test to determine the importance of terms involving *State* in the model constructed in question 6. List your hypotheses, p-value, and conclusion.
8. Fit a multiple regression model using *Year*, *Mileage*, *State*, and the interactions between *Year* and *State*, and *Mileage* and *State* to predict the *Price* of the vehicle.
9. Perform a hypothesis test to determine the importance of the terms involving *State* in the model constructed in question 8. List your hypotheses, p-value, and conclusion.

MODEL #6: Polynomial models One of the drawbacks of the linear model in homework #2 was the "free vehicle" phenomenon where the predicted price is eventually negative as the line decreases for older vehicles. Let's see if adding one or more polynomial terms might help with this. For this section you should use the dataset with vehicles from four states that you used for model 5.

10. Fit a quadratic model using *Year* to predict *Price* and examine the residuals. Construct a scatterplot of the data with the quadratic fit included. You should discuss each of the conditions for the linear model.
11. Perform a hypothesis test to determine if any of the coefficients in this model have nonzero coefficients. List your hypotheses, p-value, and conclusion.
12. You are looking at a vehicle that was 4 years old (in 2017) 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 vehicle, and include an interpretive sentence in context.
13. Does the quadratic model allow for some *Year* where a vehicle has a zero or negative predicted price? Justify your answer using a calculation or graph.
14. 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.

MODEL #7: Complete second order model For this section you should again use the dataset with vehicles from four states that you used for models 5 and 6.

15. Fit a complete second order model for predicting a used vehicle *Price* based on *Year* and *Mileage* and examine the residuals. You should discuss each of the conditions for the linear model.
16. Perform a hypothesis test to determine if any of the coefficients in this model have nonzero coefficients. List your hypotheses, p-value, and conclusion.
17. 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.
18. 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.