STOR 455 Homework #4

20 points - Due Tuesday 03/21 at 12:00pm

Theory Part

1. True or False: A variable in numbers must be quantitative.

Your answer: True

2. Suppose we included a categorical predictor with 5 categories. What is the difference in the assumptions between regressing with a coding of 1 to 5 and regressing on the 4 dummy variables created by this categorical predictor?

Your answer: When regressing with 4 dummy variables we are able to pick and choose which categories to include. This makes it so we can create separate equations for each subgroup. It also allows us to find the difference between groups by finding the difference between their equations.

Computing Part

Situation: Suppose that 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 that you get (the model) and how much it's been used. For this assignment you will investigate how the price might depend on the vehicle's year, state, and odometer reading. We focus on three states in this homework "CA", "NC", and "NY."

Data Source: To get a sample of vehicles, begin with the *UsedCars* csv file. The data was acquired by scraping Craigslist for vehicles for sale across the southeastern United States. For this assignment you will choose model of cars. Construct a subset of the *vehiclesSE* data for this model of vehicle. If your subset has cars with seemingly incorrect data (such as a price of $1, odometer reading of one million miles, year of 1900) you should remove those values from the data.

Directions: The code below should walk you through the process of selecting data from a particular model vehicle of your choice. The following R chunk begin with {r, eval=FALSE}. eval=FALSE makes this chunk not run when I knit the file. Before you knit this chunk, you should revert it to {r}.

```
library(readr)
library(car)
```

```
## Loading required package: carData
```

```
library(leaps)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v dplyr   1.0.10
## v tibble  3.1.8      v stringr 1.5.0
## v tidyr   1.3.0      v forcats 0.5.2
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

```
vehicles_all <- read_csv("UsedCars.csv", show_col_types = FALSE)

vehicles_3States = subset(vehicles_all, State=="NY"|State=="NC"|State=="CA")

# Delete the ** below and enter your chosen model
ModelOfMyChoice = "3"

vehiclesSE= subset(vehicles_3States, Model==ModelOfMyChoice)
```

Include a Categorical predictor

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

```
ThreeSeriesModel = lm(Price~State+Mileage, data = vehiclesSE)
ThreeSeriesModel
```

```
##
## Call:
## lm(formula = Price ~ State + Mileage, data = vehiclesSE)
##
## Coefficients:
## (Intercept)      StateNC      StateNY      Mileage
##  34832.3296     210.1718     428.8581      -0.2387
```

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

```
summary(ThreeSeriesModel)
```

```
##
## Call:
## lm(formula = Price ~ State + Mileage, data = vehiclesSE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21450.2  -3795.1   -649.7   3432.9  21758.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.483e+04  1.711e+02 203.565   <2e-16 ***
## StateNC      2.102e+02  2.750e+02   0.764    0.445
## StateNY      4.289e+02  2.728e+02   1.572    0.116
## Mileage     -2.387e-01  2.644e-03 -90.277   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5383 on 3036 degrees of freedom
## Multiple R-squared:  0.7324, Adjusted R-squared:  0.7322
## F-statistic:  2770 on 3 and 3036 DF,  p-value: < 2.2e-16
```

A t-test was used to see if there was a relationship between state, price, and mileage. The Ho states that p = 0 and there is no correlation between state, price, and mileage. The Ha states that p is != 0 and there is a correlation between state, price, and mileage. The p-values for the states are 0.445 and 0.116 which is significantly larger than 0.05. On the otherhand, the p-value for the mileage is less than 0.05 so that variable

is correlated with price. Overall, we fail to reject the null hypothesis and can assume the state does not influence the price.

3. Fit a multiple regression model using *Year*, *Mileage*, and *State* to predict the *Price* of the vehicle.

```
MultiModel = lm(Price~State+Mileage+Year, data = vehiclesSE)
MultiModel
```

```
##
## Call:
## lm(formula = Price ~ State + Mileage + Year, data = vehiclesSE)
##
## Coefficients:
## (Intercept)      StateNC      StateNY      Mileage         Year
##  -2.225e+06    2.542e+02    4.235e+02   -1.447e-01    1.121e+03
```

4. Perform a hypothesis test to determine the importance of terms involving *State* in the model constructed in question 3. List your hypotheses, p-value, and conclusion.

```
summary(MultiModel)
```

```
##
## Call:
## lm(formula = Price ~ State + Mileage + Year, data = vehiclesSE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15719.2  -3568.2   -908.2   3039.0  19959.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.225e+06  9.785e+04 -22.743   <2e-16 ***
## StateNC      2.542e+02  2.536e+02   1.002   0.3163
## StateNY      4.235e+02  2.516e+02   1.683   0.0924 .
## Mileage     -1.447e-01  4.745e-03 -30.493   <2e-16 ***
## Year         1.121e+03  4.851e+01  23.099   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4965 on 3035 degrees of freedom
## Multiple R-squared:  0.7724, Adjusted R-squared:  0.7721
## F-statistic:  2576 on 4 and 3035 DF,  p-value: < 2.2e-16
```

A t-test was used to see if there was a relationship between year, state, price, and mileage. The Ho states that p = 0 and there is no correlation between year, state, price, and mileage. The Ha states that p is != 0 and there is a correlation between year, state, price, and mileage. The p-values for the states are 0.3163 and 0.0924 which is larger than 0.05. On the otherhand, the p-values for mileage and year is less than 0.05 so those variables are correlated with price. Overall, we fail to reject the null hypothesis and can assume the state does not influence the price.

5. 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. Refer this model as the *Full* model.

```
Full = lm(Price~State+Mileage+Year+Year*State+Mileage*State, data = vehiclesSE)
Full
```

```
##
## Call:
```

```
## lm(formula = Price ~ State + Mileage + Year + Year * State +
##     Mileage * State, data = vehiclesSE)
##
## Coefficients:
##     (Intercept)          StateNC          StateNY          Mileage
##      -2.132e+06        7.208e+04       -7.636e+05       -1.483e-01
##            Year     StateNC:Year     StateNY:Year  StateNC:Mileage
##       1.074e+03       -3.608e+01        3.794e+02        1.305e-02
## StateNY:Mileage
##       4.340e-04
```

6. Perform a hypothesis test to determine the importance of the terms involving *State* in the model constructed in question 5. List your hypotheses, p-value, and conclusion.

```
summary(Full)
```

```
##
## Call:
## lm(formula = Price ~ State + Mileage + Year + Year * State +
##     Mileage * State, data = vehiclesSE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15945.8  -3540.1   -934.2   2968.9  20888.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.132e+06  1.215e+05 -17.555  < 2e-16 ***
## StateNC          7.208e+04  2.398e+05   0.301  0.76371
## StateNY         -7.636e+05  2.934e+05  -2.603  0.00929 **
## Mileage         -1.483e-01  5.860e-03 -25.310  < 2e-16 ***
## Year             1.074e+03  6.022e+01  17.842  < 2e-16 ***
## StateNC:Year    -3.608e+01  1.189e+02  -0.303  0.76154
## StateNY:Year     3.794e+02  1.454e+02   2.609  0.00913 **
## StateNC:Mileage  1.305e-02  1.160e-02   1.124  0.26103
## StateNY:Mileage  4.340e-04  1.468e-02   0.030  0.97641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4943 on 3031 degrees of freedom
## Multiple R-squared:  0.7747, Adjusted R-squared:  0.7742
## F-statistic:  1303 on 8 and 3031 DF,  p-value: < 2.2e-16
```

A t-test was used to see if there was a relationship between year, state, price, mileage, state and year, and state and mileage. The Ho states that p = 0 and there is no correlation between year, state, price, and mileage. The Ha states that p is != 0 and there is a correlation between year, state, price, and mileage, state and year, and state and mileage. The p-values for the states are 0.76371 and 0.00929. This means that for the state of New York it is correlated with price. For the interactions StateNC:Year, StateNC:Mileage, StateNY:Mileage these variables are all not correlated since they are above 0.05 but the StateNY:Year is below 0.05 so that is correlated. Although there is correlation with the StateNY and the StateNY:Year we fail to reject the null hypothesis and can assume the state does not influence the price.

7. Select a subset of predictors in *Full* using each of the four methods: all subsets, backward elimination, forward selection, and stepwise regression. Use Mallows' Cp (AIC) as the criterion.

```r
# All Subsets
all = regsubsets(Price~State+Mileage+Year+Year*State+Mileage*State, data = vehiclesSE)
summary(all)
```

```
## Subset selection object
## Call: regsubsets.formula(Price ~ State + Mileage + Year + Year * State +
##     Mileage * State, data = vehiclesSE)
## 8 Variables  (and intercept)
##                 Forced in Forced out
## StateNC             FALSE      FALSE
## StateNY             FALSE      FALSE
## Mileage             FALSE      FALSE
## Year                FALSE      FALSE
## StateNC:Year        FALSE      FALSE
## StateNY:Year        FALSE      FALSE
## StateNC:Mileage     FALSE      FALSE
## StateNY:Mileage     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          StateNC StateNY Mileage Year StateNC:Year StateNY:Year StateNC:Mileage
## 1  ( 1 ) " "     " "     "*"     " "  " "          " "          " "
## 2  ( 1 ) " "     " "     "*"     "*"  " "          " "          " "
## 3  ( 1 ) " "     " "     "*"     "*"  " "          " "          "*"
## 4  ( 1 ) " "     "*"     "*"     "*"  " "          "*"          " "
## 5  ( 1 ) " "     "*"     "*"     "*"  " "          "*"          "*"
## 6  ( 1 ) " "     "*"     "*"     "*"  "*"          "*"          "*"
## 7  ( 1 ) "*"     "*"     "*"     "*"  "*"          "*"          "*"
## 8  ( 1 ) "*"     "*"     "*"     "*"  "*"          "*"          "*"
##          StateNY:Mileage
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) "*"
```

```r
# Backward elimination
MSE = (summary(Full)$sigma)^2
step(Full, scale=MSE)
```

```
## Start:  AIC=9
## Price ~ State + Mileage + Year + Year * State + Mileage * State
##
##                 Df Sum of Sq        RSS      Cp
## - State:Mileage  2  31870188 7.4098e+10  6.3042
## <none>                       7.4066e+10  9.0000
## - State:Year     2 184791773 7.4250e+10 12.5623
##
## Step:  AIC=6.3
## Price ~ State + Mileage + Year + State:Year
##
##              Df  Sum of Sq        RSS       Cp
## <none>                    7.4098e+10   6.3042
```

```
## - State:Year  2 7.2682e+08 7.4824e+10  32.0480
## - Mileage     1 2.3028e+10 9.7126e+10 946.6865
```

```
##
## Call:
## lm(formula = Price ~ State + Mileage + Year + State:Year, data = vehiclesSE)
##
## Coefficients:
##  (Intercept)       StateNC       StateNY       Mileage          Year
##   -2.186e+06     2.956e+05    -7.509e+05    -1.453e-01     1.101e+03
## StateNC:Year  StateNY:Year
##   -1.468e+02     3.732e+02
```

```
# Forward selection
none = lm(Price~1, data = vehiclesSE)
step(none, scope = list(upper=Full), scale = MSE, direction = "forward")
```

```
## Start:  AIC=10418.07
## Price ~ 1
##
##           Df  Sum of Sq        RSS       Cp
## + Mileage  1 2.4076e+11 8.8058e+10   567.59
## + Year     1 2.3088e+11 9.7933e+10   971.71
## + State    2 4.6596e+09 3.2415e+11 10231.39
## <none>                  3.2881e+11 10418.07
##
## Step:  AIC=567.59
## Price ~ Mileage
##
##         Df  Sum of Sq        RSS      Cp
## + Year   1 1.3151e+10 7.4907e+10  31.419
## <none>                8.8058e+10 567.592
## + State  2 7.8604e+07 8.7979e+10 568.375
##
## Step:  AIC=31.42
## Price ~ Mileage + Year
##
##         Df Sum of Sq        RSS     Cp
## <none>               7.4907e+10 31.419
## + State  2  82363970 7.4824e+10 32.048
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Year, data = vehiclesSE)
##
## Coefficients:
## (Intercept)      Mileage         Year
##  -2.225e+06   -1.447e-01    1.120e+03
```

```
# Stepwise regression
step(none, scope=list(upper=Full), scale = MSE)
```

```
## Start:  AIC=10418.07
## Price ~ 1
##
##           Df  Sum of Sq        RSS       Cp
## + Mileage  1 2.4076e+11 8.8058e+10   567.59
```

```
## + Year     1 2.3088e+11 9.7933e+10   971.71
## + State    2 4.6596e+09 3.2415e+11 10231.39
## <none>                  3.2881e+11 10418.07
##
## Step:  AIC=567.59
## Price ~ Mileage
##
##           Df  Sum of Sq        RSS        Cp
## + Year     1 1.3151e+10 7.4907e+10    31.419
## <none>                  8.8058e+10   567.592
## + State    2 7.8604e+07 8.7979e+10   568.375
## - Mileage  1 2.4076e+11 3.2881e+11 10418.073
##
## Step:  AIC=31.42
## Price ~ Mileage + Year
##
##           Df  Sum of Sq        RSS      Cp
## <none>                  7.4907e+10  31.419
## + State    2 8.2364e+07 7.4824e+10  32.048
## - Year     1 1.3151e+10 8.8058e+10 567.592
## - Mileage  1 2.3026e+10 9.7933e+10 971.710
##
## Call:
## lm(formula = Price ~ Mileage + Year, data = vehiclesSE)
##
## Coefficients:
## (Intercept)      Mileage         Year
##  -2.225e+06   -1.447e-01    1.120e+03
```

When comparing the AIC of these models the best is Price ~ Mileage + Year + State + State:Year which has the lowest AIC value of 6.3.

8. Assess and compare the overall effectiveness of the four models (some or all of them may be identical).

The most effective model is the Price ~ Mileage + Year + State + State:Year model. The next closest is Price~State + Mileage + Year + State:Year + State:Mileage with an AIC of 9. After that it drops off heavily and are much less of a fit. Price ~ Mileage + Year with an AIC of 31.42, Price~Mileage with an AIC of 567.59, and Price~1 with an AIC of 10418.07. Overall, Price ~ Mileage + Year + State + State:Year is quite a good fit due to it's very low AIC value especially when compared to the other models.