STOR 455 Homework #6

20 points - Due Tuesday 10/24 at 12:30pm

Theory Part

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

Your answer: False. For example, zip code is a number, but it is not a quantitative variable

2. True or False: Using all subsets selection over 6 variables will result in a consideration of 64 models.

Your answer: False, this is because it will result in 63 models since the formula is 2^k-1 or 2^6-1 = 63

3. If $n = 100, k = 3$, $SSE = 100$, and $SST = 200$, what is the adjusted R^2?

```
Rsqrd <-(100/(100-3-1))/(200/(100-1))
1-Rsqrd
```

```
## [1] 0.484375
```

Computing Part

Instructions: You may (and should) collaborate with other students. However, you must complete the assignment by yourself. You should complete this assignment in an R Notebook, including all calculations, plots, and explanations. Make use of the white space outside of the R chunks for your explanations rather than using comments inside of the chunks. For your submission, you should knit the notebook to PDF (it is usually smoother first knit to Word then save the file as pdf) and submit the file to Gradescope. The submitted PDF should not be longer than 20 pages.

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 two states in this homework "CA" and "NC".

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}.**

```r
library(readr)

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

vehicles_2States = subset(vehicles_all, State=="NC"|State=="CA")

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


vehiclesSE= subset(vehicles_2States, Model==ModelOfMyChoice)
```

Questions

Q1

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

```
#
Mod1 = lm(Price~Mileage+State, data = vehiclesSE)
summary(Mod1)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + State, data = vehiclesSE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9526.7 -1835.5  -277.9  1735.5 15065.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.121e+04  1.015e+02 208.912  < 2e-16 ***
## Mileage     -9.891e-02  1.394e-03 -70.943  < 2e-16 ***
## StateNC      6.192e+02  1.237e+02   5.007 5.95e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2794 on 2298 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6863
## F-statistic:  2517 on 2 and 2298 DF,  p-value: < 2.2e-16
```

Q2

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.

Hypotheses: –Null Hypothesis (H0): The "State" Coefficients is zero for all state in {1,2} –Alternative Hypothesis (H1): The "State" Coefficient is nonzero for all state in {1,2}

Conclusions: P-Value: 5.954393e-07 –Reject the null. There is statistically significant evidence (p-vale=5.954393e-07) to suggest that the at least one state coefficient is nonzero.

```
#
mod2 = lm(Price~Mileage, data = vehiclesSE)
anovamod <- anova(mod2,Mod1)
anovamod
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Mileage
## Model 2: Price ~ Mileage + State
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1   2299 1.8139e+10
## 2   2298 1.7943e+10  1 195736940 25.068 5.954e-07 ***
```

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

```
pvalmod<-anovamod$"Pr(>F)"[1:2]
pvalmod
```

```
## [1]           NA 5.954393e-07
```

Q3

Fit a multiple regression model using *Year*, *Mileage*, *State*, as well as two interaction terms of *State* and *Year* and *Mileage* to predict the *Price* of the vehicle.

```
#
mod3 <- lm(Price~Year+Mileage+State+ Year*State + Mileage*State, data = vehiclesSE)
summary(mod3)
```

```
##
## Call:
## lm(formula = Price ~ Year + Mileage + State + Year * State +
##     Mileage * State, data = vehiclesSE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5674.6 -1649.0  -274.9  1341.9 10235.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.756e+06  6.734e+04 -26.078   <2e-16 ***
## Year             8.813e+02  3.339e+01  26.394   <2e-16 ***
## Mileage         -4.532e-02  2.614e-03 -17.340   <2e-16 ***
## StateNC          1.343e+05  1.055e+05   1.272    0.203
## Year:StateNC    -6.637e+01  5.233e+01  -1.268    0.205
## Mileage:StateNC -9.589e-04  3.984e-03  -0.241    0.810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2296 on 2295 degrees of freedom
## Multiple R-squared:  0.7888, Adjusted R-squared:  0.7883
## F-statistic:  1714 on 5 and 2295 DF,  p-value: < 2.2e-16
```

Q4

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.

Hypotheses: –Null Hypothesis (H0): The "State" Coefficients is zero for all state in {2,3,4,5} –Alternative Hypothesis (H1): The "State" Coefficient is nonzero for all state in {2,3,4,5}

Conclusions: P-value: 1.358142e-08 –Reject the null. There is statistically significant evidence (p-vale=1.358142e-08) to suggest that the at least one state coefficient is nonzero.

```r
#
mod4<- lm(Price~Year+Mileage, data = vehiclesSE)
anovamod2<- anova(mod4, mod3)
anovamod2
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Year + Mileage
## Model 2: Price ~ Year + Mileage + State + Year * State + Mileage * State
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1   2298 1.2303e+10
## 2   2295 1.2093e+10  3 209914345 13.279 1.358e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
pvalmod2<- anovamod2$'Pr(>F)'[1:2]
pvalmod2
```

```
## [1]           NA 1.358142e-08
```