STOR 455 Homework #7

20 points - Due Tuesday 04/27 at 12:00pm

Theory Part

Below is a balanced two-way ANOVA table with the consideration of interaction effect. In this analysis there are 16 observations. There are two factors, each with two levels. Complete this table by filling in missing values.

	Df	Sum Sq	Mean Sq	F value
FactorA	1	124609	124609	52.63
FactorB	1	252004	252004	106.44
FactorA:FactorB	1	51307.41	51307.41	21.67
Residuals	12	28412	2367.67	None

Computing Part

Situation (again): Suppose that 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, mileage, and the model of car.

Data Source: Your sample of cars will again be taken from the UsedCar CSV file on Sakai. 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 will need to select six new samples, each with *exactly* 50 vehicles, for six different *Model* of used vehicles for sale in North Carolina from the UsedCar dataset. There will likely be more than 50 of your selected models for sale in North Carolina, so you should randomly select those 50 vehicles from the larger number that are available. The six models of vehicles should be selected such that three models of vehicles are selected from Japanese companies, and another three from US companies (i.e. *Make*; It does not matter where the cars were actually manufactured). Within each country, you should select a compact car, a mid-sized car, and a SUV (Note that the country and types of vehicles are not given in the data and are for you to determine). You should add new variables to the dataframes for the country of the company and type of vehicle (compact vs mid-sized vs SUV) and combine these six samples into one dataframe (use the function `rbind(sample1,sample2,...,sample6)` here to combine samples). When selecting these samples make sure to use `set.seed()`. This will select the same sample each time that you run (and knit) your code.The code below is an example of how you could select a random sample of 50 cars for a given model:

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# This line will only run if the UsedCars.csv is stored in the same directory as this notebook!
UsedCars <- read_csv("UsedCars.csv")
```

```
## Rows: 1048575 Columns: 9
```

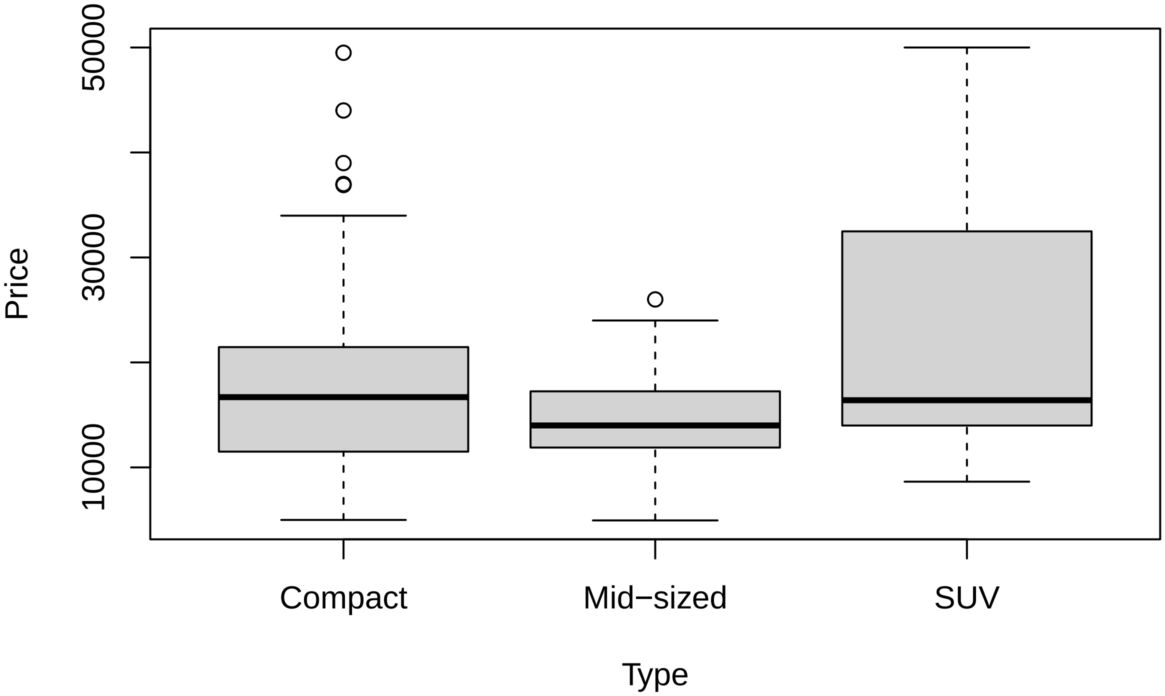
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): City, State, Vin, Make, Model
## dbl (4): Id, Price, Year, Mileage
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Creates a dataframe with the number of each model for sale in North Carolina

# Sample code to select 50 Civic sold in NC from the Usedcars dataset. You can choose a different count
set.seed(1)
Civic = sample_n(subset(UsedCars, Make=="Honda" & State=="NC" & Model=="Civic"), 50)
Accord = sample_n(subset(UsedCars, Make=="Honda" & State=="NC" & Model=="Accord"), 50)
Rogue = sample_n(subset(UsedCars, Make=="Nissan" & State=="NC" & Model=="Rogue"), 50)

CTS = sample_n(subset(UsedCars, Make=="Cadillac" & State=="NC" & Model=="CTS"), 50)
Limited = sample_n(subset(UsedCars, Make=="Chrysler" & State=="NC" & Model== "200Limited"), 50)
Expedition = sample_n(subset(UsedCars, Make=="Ford" & State=="NC" & Model=="Expedition"), 50)

New = rbind(Civic, Accord, Rogue, CTS, Limited, Expedition)

New$Country = ifelse(New$Model=="Civic"|New$Model=="Accord"|New$Model=="Rogue","Japan","USA")
New$Type = ifelse(New$Model=="Civic"|New$Model=="CTS", "Compact", ifelse(New$Model=="Accord"|New$Model==
```

One Way ANOVA

1. Produce a set of side-by-side boxplots to compare the price distributions of your three types of vehicles (not the models). Comment on any obvious differences in the distributions.

```
boxplot(Price~Type, data=New)
```



```
tapply(New$Price,New$Type, mean)
```

```
##   Compact Mid-sized       SUV
##  17776.63  14488.57  22953.01
```

```
tapply(New$Price,New$Type, sd)
```

```
##   Compact Mid-sized       SUV
##  8599.652  4212.270 11208.139
```

The medians for all types of cars are similiar. The min for the Compact and Mid-sized is about the same while the SUV min is higher. The Q3-Median for the SUV is significantly larger compared to the Median-Min. The Max for the SUV is significantly higher than the maxes of the Mid-Sized and Compact cars. The compact car has 4 outliers above the max, the Mid-sized has 1, and the SUV has no outliers.

2. Produce summary statistics (mean and standard deviation) for each of the groups (vehicle types) AND the entire sample of vehicle prices.

```
tapply(New$Price,New$Type, mean)
```

```
##   Compact Mid-sized       SUV
##  17776.63  14488.57  22953.01
```

```
tapply(New$Price,New$Type, sd)
```

```
##   Compact Mid-sized       SUV
##  8599.652  4212.270 11208.139
```

```
mean(New$Price)
```

```
## [1] 18406.07
```

```
sd(New$Price)
```

```
## [1] 9172.521
```

3. Construct an ANOVA model for the mean price by vehicle type. Include the output showing the ANOVA table; state hypotheses, and provide a conclusion in the context of your data.

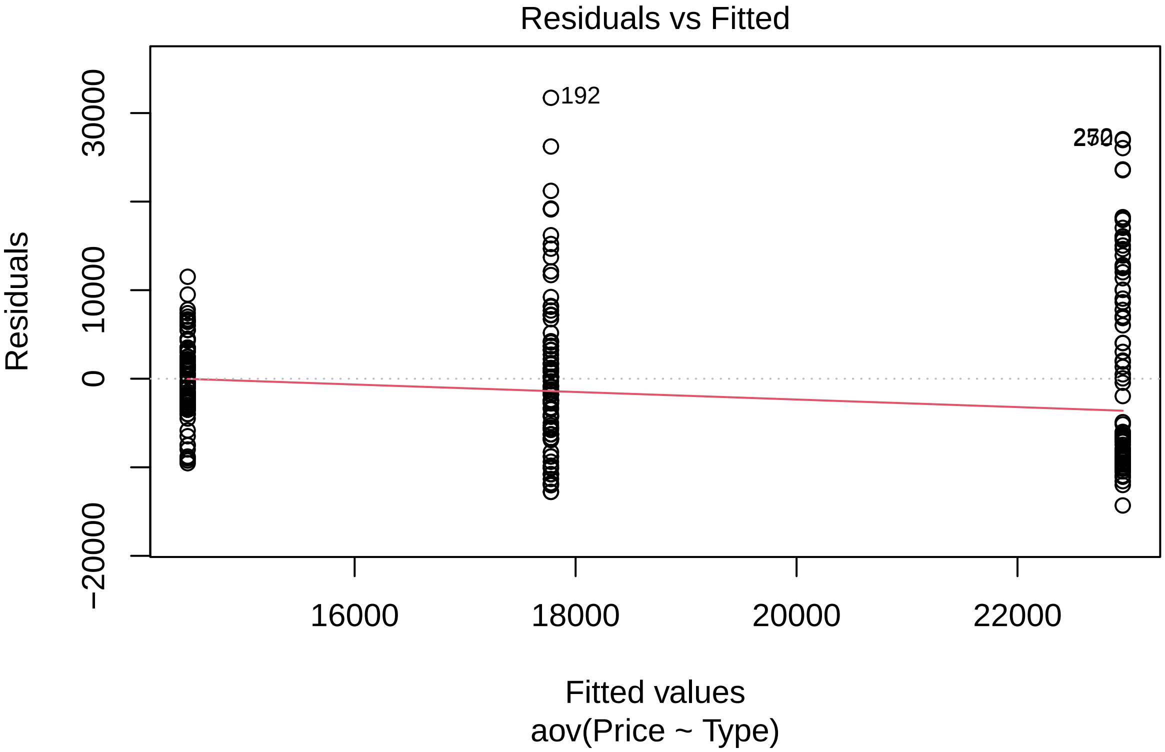
```
AnovaModel = aov(Price~Type,data = New)
summary(AnovaModel)
```

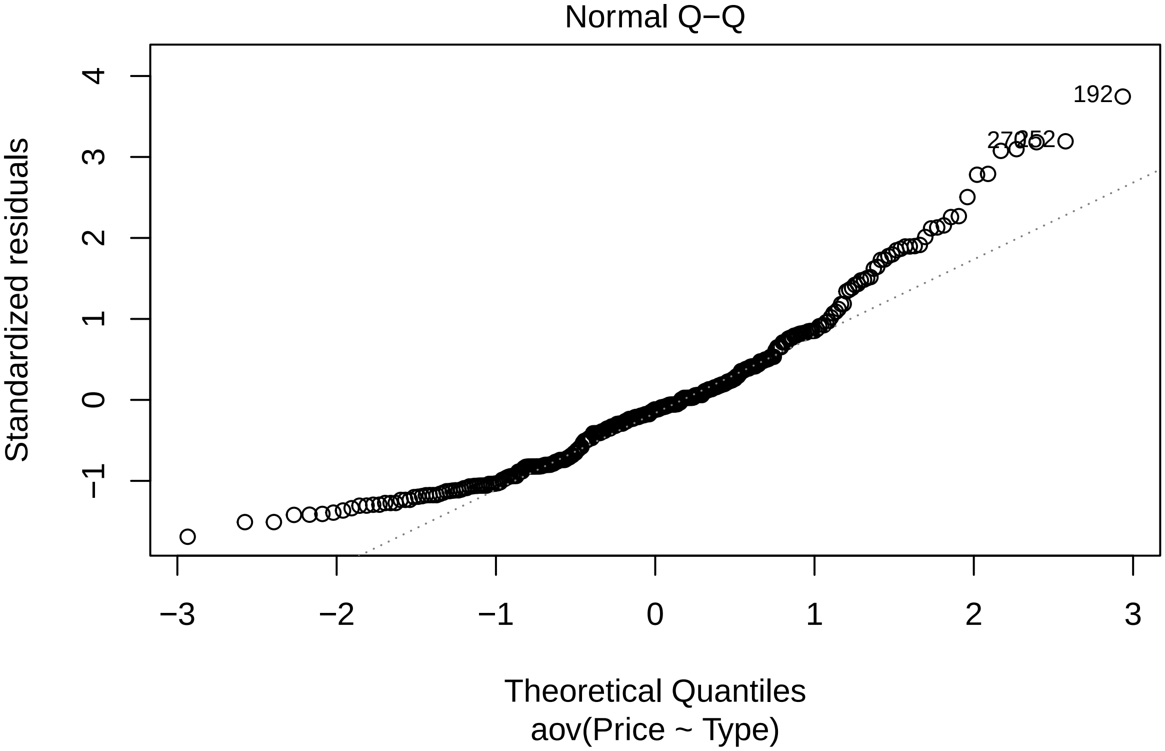
```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)    
## Type          2 3.642e+09 1.821e+09   25.14 8.22e-11 ***
## Residuals   297 2.151e+10 7.244e+07                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An F test was used to see if there was a relationship between Price and Type. The Ho states that p = 0 and there is no correlation between Price and Type. The Ha states that p is != 0 and there is a correlation between Price and Type. The p-value for Type is 8.22eˆ-11 which is greater than 0. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Type.

4. Produce plots and/or summary statistics to comment on the appropriateness of the following conditions for your data: normality of the residuals, and equality of the variances. If you find that the conditions are *not* met, You can still continue with analysis of your data for this homework. We will soon discuss how to deal with violations of these conditions.

```
plot(AnovaModel, 1:2)
```





When looking at the residual plots all the residuals are all relatively normal and centered around 0 with the Compact and Mid-sized more so compared to the SUVs. For the Compact and the SUVs the variance is about the same with it being relatively spread out but the variance for the Mid-sized cars is quite small. The qq plot is also relatively linear meaning that the model meets the conditions. Overall, I would say the conditions are met as all the car Types are normal and have similar variances.

5. If your ANOVA model indicates that there are significant differences among the vehicle type price

means, discuss where the significant differences occur using Tukey HSD methods. If your ANOVA indicates there are not significant differences among the vehicle type price means, determine how different your means prices would need to be in order to find a significant difference using the Tukey HSD methods.

```
TukeyHSD(AnovaModel)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price ~ Type, data = New)
##
## $Type
##                       diff       lwr       upr     p adj
## Mid-sized-Compact -3288.06 -6123.311  -452.809 0.0182755
## SUV-Compact        5176.38  2341.129  8011.631 0.0000686
## SUV-Mid-sized      8464.44  5629.189 11299.691 0.0000000
```

Comparing the Mid-size-Compact vs the other two comparisons there is a significant difference. This is due to the Compact's mean and SD being much smaller compared to the mean/sd of the Mid-sized cars and the overall mean/sd.

Two Way ANOVA

6. Construct an ANOVA model for the mean price using the country of the company and the type of vehicle as predictors (without an interaction). Include the output showing the ANOVA table; state hypotheses and provide a conclusion in the context of your data. If your ANOVA model indicates there are significant differences among the vehicle price means: Discuss where the significant differences occur using Tukey HSD methods.

```
TwoWayModel = aov(Price~Type+Country,data = New)
summary(TwoWayModel)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## Type          2 3.642e+09 1.821e+09   32.74 1.43e-13 ***
## Country       1 5.050e+09 5.050e+09   90.79  < 2e-16 ***
## Residuals   296 1.646e+10 5.562e+07
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

An F test was used to see if there was a relationship between Price and Type. The Ho states that p = 0 and there is no correlation between Price and Type. The Ha states that p is != 0 and there is a correlation between Price and Type. The p-value for Type is 1.43e^-13 which is greater than 0. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Type.

An F test was used to see if there was a relationship between Price and Country. The Ho states that p = 0 and there is no correlation between Price and Country. The Ha states that p is != 0 and there is a correlation between Price and Country. The p-value for Type is 2e^-16 which is greater than 0. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Country.

7. Construct an ANOVA model for the mean price using the country of the company and the type of vehicle as predictors with the interaction. Include the output showing the ANOVA table; state

hypotheses and provide a conclusion in the context of your data. If your ANOVA indicates that there are significant differences among the car price means: Discuss where the significant differences occur using Tukey HSD methods.

```
InteractModel = aov(Price~Type+Country+Type*Country,data = New)
summary(InteractModel)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)    
## Type            2 3.642e+09 1.821e+09   43.64 <2e-16 ***
## Country         1 5.050e+09 5.050e+09  121.04 <2e-16 ***
## Type:Country    2 4.198e+09 2.099e+09   50.31 <2e-16 ***
## Residuals     294 1.227e+10 4.172e+07                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An F test was used to see if there was a relationship between Price and Type. The Ho states that p = 0 and there is no correlation between Price and Type. The Ha states that p is != 0 and there is a correlation between Price and Type. The p-value for Type is 2eˆ-16 which is greater than 0. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Type.

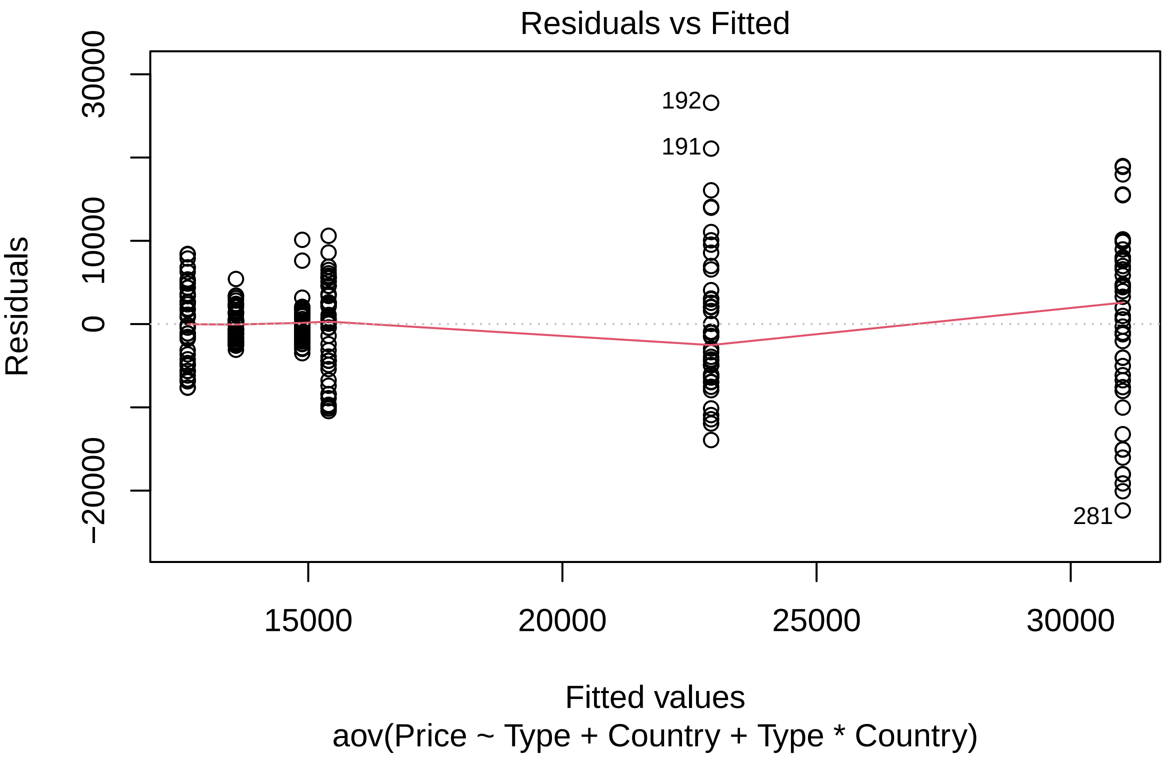
An F test was used to see if there was a relationship between Price and Country. The Ho states that p = 0 and there is no correlation between Price and Country. The Ha states that p is != 0 and there is a correlation between Price and Country. The p-value for Type is 2eˆ-16 which is greater than 0. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Country.

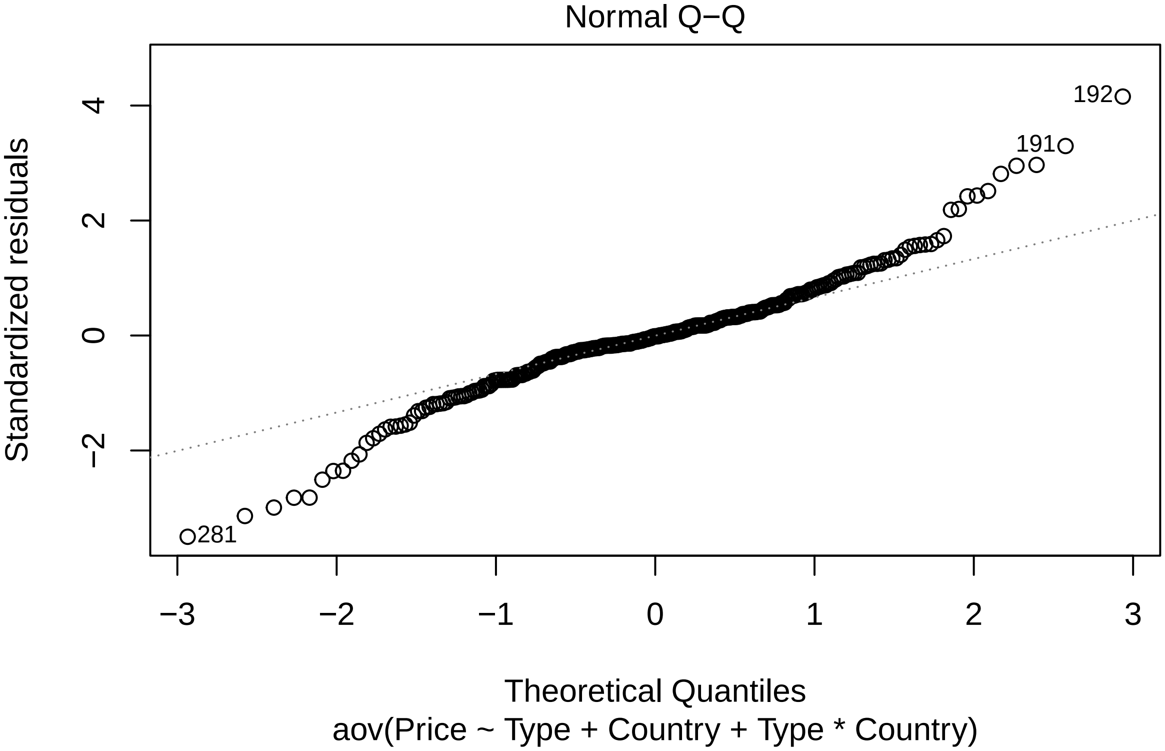
An F test was used to see if there was a relationship between Price and Type times Country. The Ho states that p = 0 and there is no correlation between Price and Country. The Ha states that p is != 0 and there is a correlation between Price and Type times Country. The p-value for Type times Country is 2eˆ-16 which is greater than 0. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Type times Country.

With all of the p-values being 2eˆ-16 there is no significant difference between the car price means. This means that car type and country are valid in predicting the price of a car.

8. Produce two interaction plots for the previous model. If you found significant interactions in your hypothesis test, comment on how these interactions are shown in the plot. If you did not find significant interactions in your hypothesis test, comment on how the (lack of) interactions are shown in the plot.

```
plot(InteractModel, 1:2)
```





When looking at the Residuals plot all the values are centered around or near 0. For all the values the variance isn't really the same with the left side more clustered around 0 compared to the right side where the values are more spread out. When looking at the qq plot the model is relatively linear meaning that the model meets the conditions. Overall, with a slight exception to the end values I would say the conditions are met as the model looks normal and have somewhat similar variances.