Midterm Exam - Open Book Section (R) - Part 2 Solutions

Recommended Packages

```
# Load relevant libraries (add here if needed)
library(car)
```

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

Car Purchasing Data Analysis

For this exam, you will be building a model to predict car purchase prices (*Car.Purchase.Amount*) that are sold in different countries of Mexico, Canada, and USA.

The "Car_Purchasing_Data.csv" data set consists of the following variables:

- *Country*: country in which the car is sold (3-letter identifier)
- *Gender*: gender of the buyer (0=female, 1=male)
- *Age*: age of the buyer (years)
- *Annual.Salary*: annual salary earned of the buyer ($ USD)
- *Credit.Card.Debt*: amount of reported credit debt owed by buyer ($ USD)
- *Net.Worth*: amount of reported assets of buyer ($ USD)
- *Car.Purchase.Amount*: amount the car was purchased for by buyer ($ USD)

Read the data and answer the questions below. Assume a significance threshold of 0.05 for hypothesis tests unless stated otherwise.

```
# Read the data set
buyers = read.csv('Car_Purchasing_Data.csv', header=TRUE)

#Set Gender & Country as a categorical variable
buyers$Gender<-as.factor(buyers$Gender)
buyers$Country<-as.factor(buyers$Country)

#View first few rows
head(buyers)
```

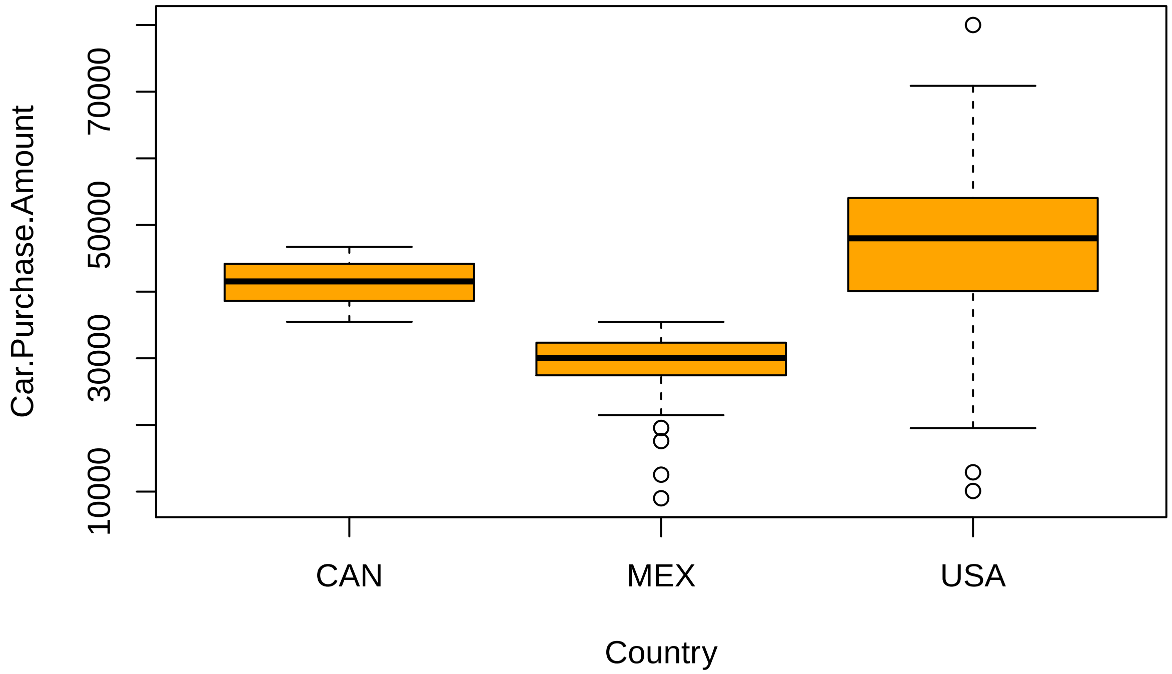
```
##   Country Gender Age Annual.Salary Credit.Card.Debt Net.Worth
## 1     USA      1  55      83333.81         9874.075 1000000.0
## 2     USA      0  48      86565.16        13701.800  819002.2
## 3     USA      1  62      66655.41         8001.644  805075.5
## 4     USA      0  60      81565.96         9072.063  544292.0
## 5     USA      0  55      70787.28        10155.341  853913.9
## 6     USA      1  61      79792.13        14245.533  497950.3
##   Car.Purchase.Amount
## 1            80000.00
## 2            70878.30
## 3            70598.97
## 4            69669.47
## 5            68925.09
## 6            68678.44
```

Note: For all of the following questions, treat all variables as quantitative variables except for *Gender* and *Country*. They have already been converted to categorical variables in the above code.

Question 1 - Exploratory Data Analysis of Categorical Variable - 2pts

Create a boxplot of the response variable *Car.Purchase.Amount* and the categorical variable *Country*. From this plot, does *Country* appear useful in predicting *Car.Purchase.Amount*? Explain how you came to your conclusion.

```
#code for boxplot...
boxplot(Car.Purchase.Amount~Country, data=buyers, col="orange")
```



Response to Question 1: Yes, *Country* appears useful in predicting *Car.Purchase.Amount*. All three medians appear to be vastly different. In addition, we see the interquantile ranges (orange boxes) do not all share the same values. Thus, there is a good chance the predicting variable *Country* will provide predictive power.

Question 2 - ANOVA - 4pts

Create an ANOVA model called *anovamodel* to compare the mean *Car.Purchase.Amount* among the different countries. Display the corresponding ANOVA table.

A) Identify the value of the sum of square treatment (SSTr) from the ANOVA table.

B) Provide the variable formula that is used to calculate the F-Value in the table.

```
#Code to create ANOVA model...

anovamodel<-aov(Car.Purchase.Amount~Country, data=buyers)
summary(anovamodel)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Country       2 1.541e+10 7.704e+09   90.08 <2e-16 ***
## Residuals   497 4.251e+10 8.553e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Response to Question 2A: SSTr = 1.541e+10

Response to Question 2B: MSSTr/MSE, or Ratio between Mean Sum of Squares Treatment to Mean Sum of Squares Errors, or Mean Sq Country/Mean Sq Residuals, or 7.704e09/8.553e07, etc.

Note: Variables are acceptable or numerical values for each variable are acceptable

Question 3 Test of Equal Means - 4pts

A) State the Null and Alternative Hypotheses for the Test of Equal Means.

B) Do you reject or fail to reject the null hypothesis for the test of equal means at a significant level of 0.05? Explain your answer using the values from the ANOVA Table

C) Given your answer in part B, explain what this conclusion means in the context of the problem.

Response to Question 3A:

H0: $\mu_1 = \mu_2 = \mu_3$

HA: At least one pair of means are not statistically equal

Response to Question 3B: Since the p-value(<2e-16) < alpha(0.05), we REJECT the null hypothesis.

Response to Question 3C: At least one pair of means is not equal. We will need to perform further analysis to identify which pair(s) is/are not equal.

Question 4 Pairwise Comparison - 4pts

Conduct a pairwise comparison of the mean *Car.Purchase.Amount* for the different categories of *Country*. Use a 95% confidence level for this comparison.

A) According to the pairwise comparison, identify all the pairs that are statistically significantly different.

B) State how you came to your conclusion.

C) Provide an interpretation of “diff” in the context of the average purchase price between a car sold in Mexico and Canada. (Note: provide this interpretation regardless of the means being statistically different/equal).

```
# Code to create pairwise-comparison...
TukeyHSD(anovamodel)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Car.Purchase.Amount ~ Country, data = buyers)
##
## $Country
##               diff       lwr       upr p adj
## MEX-CAN -12614.241 -16407.76 -8820.727 0e+00
## USA-CAN   5827.148   3397.63  8256.666 1e-07
## USA-MEX  18441.389  15094.09 21788.689 0e+00
```

Response to Question 4A: All the pairs that are statistically significantly different.

Response to Question 4B: We should not see the value 0 included in the confidence intervals or we can compare the pvalue to our alpha value (0.05). All intervals do not contain 0 and all pvalues < 0.05.

Response to Question 4C: On average, a car in Mexico is $12614 less than a car sold in Canada.

Question 5 Exploratory Data Analysis Quantitative Variables - 4pts

Now consider the quantitative variables ONLY: *Age*, *Annual.Salary*, *Credit.Card.Debt*, and *Net.Worth*.

Compute the correlation coefficients between each quantitative variables and also the response variable.

A) Which predicting variable has the best correlation with the *response*?

B) Interpret the value of the best correlation coefficient in the context of the problem. Include strength (weak, moderate, strong) and direction (positive, negative).

C) Considering the predicting variables, does the correlation matrix suggest signs of multicollinearity? Explain how you came to your conclusion.

```
# Code to calculate correlation...
round(cor(buyers[c(3,4,5,6,7)]),2)
```

```
##                      Age Annual.Salary Credit.Card.Debt Net.Worth
## Age                 1.00          0.00             0.03      0.02
## Annual.Salary       0.00          1.00             0.05      0.01
## Credit.Card.Debt    0.03          0.05             1.00     -0.05
## Net.Worth           0.02          0.01            -0.05      1.00
## Car.Purchase.Amount 0.63          0.62             0.03      0.49
##                     Car.Purchase.Amount
## Age                                0.63
## Annual.Salary                      0.62
## Credit.Card.Debt                   0.03
## Net.Worth                          0.49
## Car.Purchase.Amount                1.00
```

Response to Question 5A: Age has the strongest correlation with Car.Purchase.Amount with a positive correlation of 0.63.
Response to Question 5B: Age has a moderately positive relationship with the Car purchase amount.

Response to Question 5C: No, the correlation matrix does not suggest signs of multicollinearity, since all correlations between the predictors are small.

Question 6 Multiple Linear Regression - 4pts

Create a full model with **ALL** variables (quantitative and qualitative) called **lm.full** with *Car.Purchase.Amount* as the response variable. Include all variables in the dataset. Display the summary table for the model. *Note: Treat all variables as quantitative variables except for Gender and Country. Include an intercept.*

A) Which coefficients are significant at the 0.05 significance level?

B) Is the model significant overall using an alpha of 0.05? Why/Why not?

```
# Code to create model...
lm.full<-lm(Car.Purchase.Amount~., data=buyers)
summary(lm.full)
```

```
##
## Call:
## lm(formula = Car.Purchase.Amount ~ ., data = buyers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -471.9 -202.1   17.2  205.4  463.6
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      -4.213e+04  1.076e+02 -391.696   <2e-16 ***
## CountryMEX        7.206e+01  4.495e+01    1.603   0.1095
## CountryUSA        5.085e+01  2.807e+01    1.812   0.0706 .
## Gender1           2.008e+01  2.184e+01    0.919   0.3583
## Age               8.397e+02  1.440e+00  583.344   <2e-16 ***
## Annual.Salary     5.623e-01  1.005e-03  559.649   <2e-16 ***
## Credit.Card.Debt  5.722e-03  3.106e-03    1.842   0.0660 .
## Net.Worth         2.893e-02  6.429e-05  449.993   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 241.1 on 492 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 1.422e+05 on 7 and 492 DF,  p-value: < 2.2e-16
```

Response to Question 6A: Intercept, Age, Annual.Salary, Net.Worth

Response to Question 6B: Yes, the p-value of the model is 2.2e-16 which is less than the critical alpha of 0.05.

Question 7 Confidence Intervals - 3pts

What are the bounds for a **99%** confidence interval on the coefficient for *Credit.Card.Debt*? Using this confidence interval, is the coefficient for *Credit.Card.Debt* plausibly equal to zero at this confidence level? Explain.

```
# Code to calculate 99% CI...
confint(lm.full, level = 0.99)[7,]
```

```
##       0.5 %      99.5 %
## -0.00230977  0.01375308
```

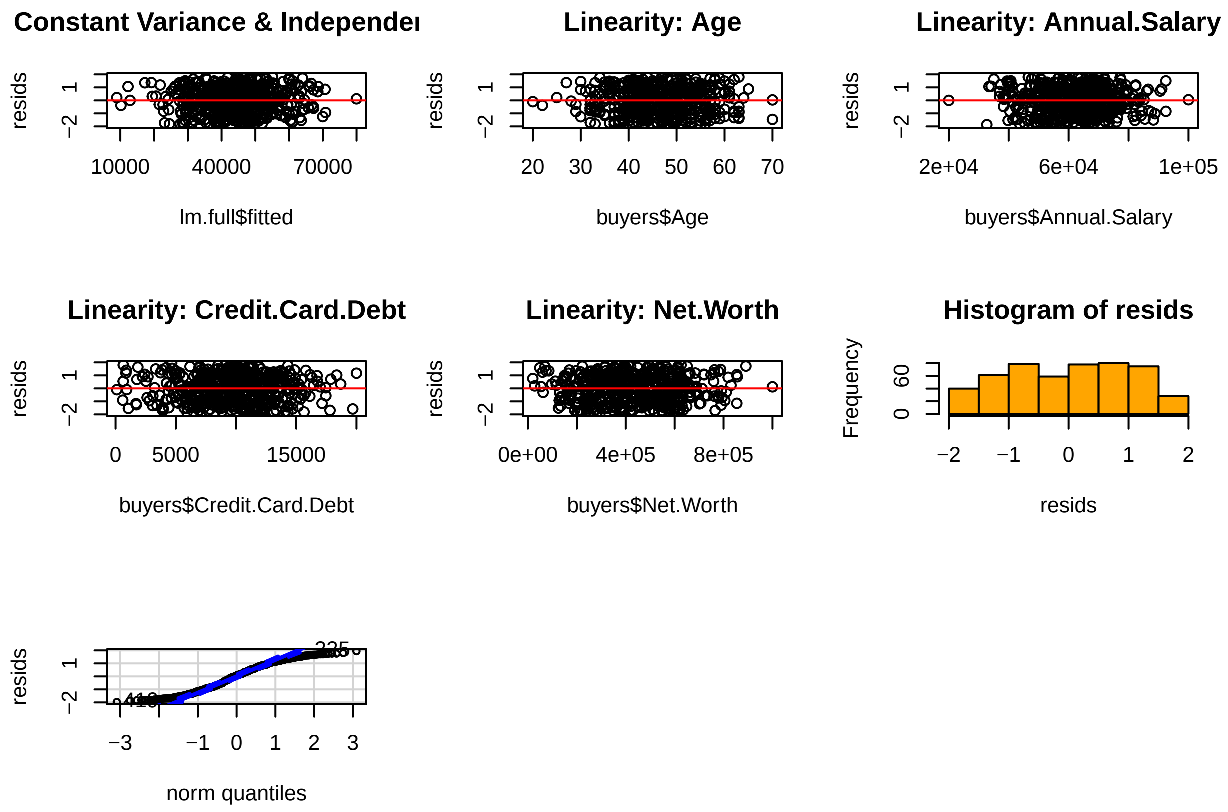
Response to Question 7: The confidence interval for the coefficient of *Credit.Card.Debt* is [-0.00230977, 0.01375308]. At the 99% confidence level *Credit.Card.Debt* is plausibly equal to zero because the value 0 is contained within the confidence interval.

Question 8 Residual Analysis - 10pts

Perform residual analysis on the *lm.full* model for the 4 assumptions. State whether the assumption holds and why you came to the conclusion.

```
#code for plots...
par(mfrow=c(3,3))

resids = rstandard(lm.full)

#Constant Variance and Independence/Correlation
plot(lm.full$fitted, resids, main = "Constant Variance & Independence")
abline(h=0, col="red")

#Linearity
plot(buyers$Age, resids, main = "Linearity: Age")
abline(h=0, col="red")
plot(buyers$Annual.Salary, resids, main = "Linearity: Annual.Salary")
abline(h=0, col="red")
plot(buyers$Credit.Card.Debt, resids, main = "Linearity: Credit.Card.Debt")
abline(h=0, col="red")
plot(buyers$Net.Worth, resids, main = "Linearity: Net.Worth")
abline(h=0, col="red")

#Normality
hist(resids, col="orange")
qqPlot(resids)
```

```
## [1] 419 225
```



Response to Constant Variance Assumption: The constant variance assumption appears to hold as the variance of the standardized residuals seems to be constant across all fitted values.

Response to Independence Assumption: We do not see any clustering in the Fitted vs Residual plot. We see an even scattering of residuals across all fitted values. Hence, the residuals appear to be uncorrelated.

Response to Linearity Assumption: The linearity/mean zero assumption appears to hold. In each predicting variable vs residual plot, we do not see any patterns. We see an even spread of residuals around the zero line through all values of the predicting variables. There are an equal number of residuals above and below the zero line across all predicting values.

Note: Any other valid plot (e.g. fitted vs residual, scatterplots of the response vs. predicting variables) is acceptable for full credit

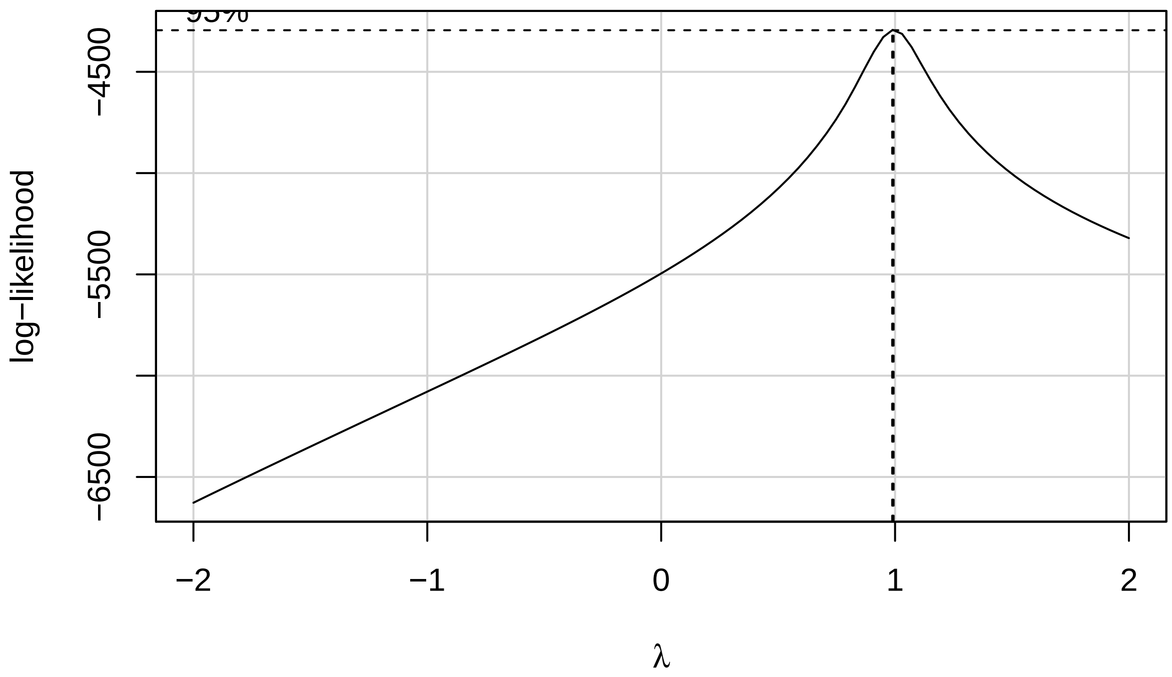
Response to Normality Assumption: The normality assumption may not hold. We do not see our normal bell shape curve in the histogram. We also see the residual fall outside the 95% quantile of the QQ Plot.

Note: One of the two plots (qq-plot and histogram) is sufficient for full credit

Question 9 Transformations - 3pts

Perform a BoxCox analysis to find if a transformation of the response variable is appropriate. State the optimal lambda value rounded to the nearest half integer and conclusion. *(Do not apply any transformations to the model)*

```
#code for boxcox...
bc<-boxCox(lm.full)
```



```
round(bc$x[which(bc$y==max(bc$y))])
```

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

Response to Question 9:

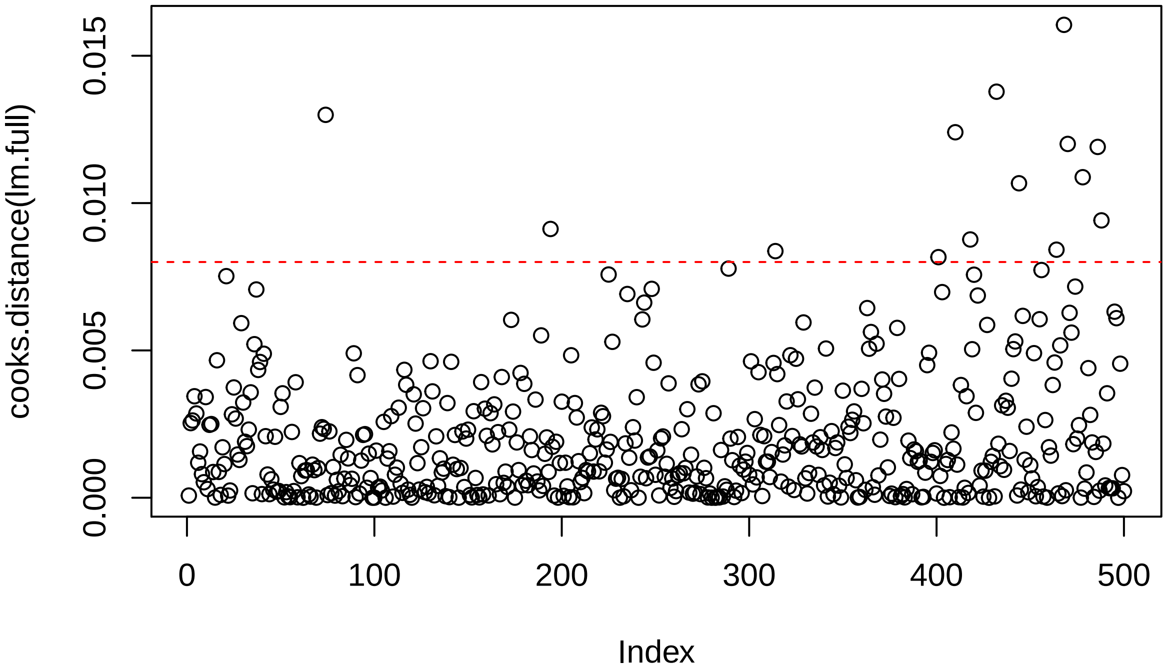
Optimal lambda value:: 0.989899 or 1

Conclusion: Using the BoxCox Transformation, a lambda value of 1 is the optimal transformation which equates to no transformation.

Question 10 - Outlier Detection - 3pts

Using Cook's distances, evaluate whether there are any outliers in the *lm.full* model. Display your plot and state your conclusion. *(Do not remove any observations)*

```
#code for outlier detection...
plot(cooks.distance(lm.full))
abline(h=4/nrow(buyers), col="red", lty=2)
```



Response to Question 11: There are not distinctively high values compared to the rest, with the highest Cook's distance being less than 0.02. Using a threshold of 1, we observe that there is not any such Cook's distance that is greater than 1 in our data. Using the threshold of 4/n (red line), we observe that there are several points that have a Cook's distance greater than this threshold; however, this does not necessarily indicate that they are all outliers but most likely that they are the tail of a heavy-tailed distribution, as seen in Q8.

Hence, it is reasonable to say that based on Cook's distance, outliers are not a concern in this dataset.

Question 11 - Reduced Model 6pts

Create a third model called *lm.red* by removing *credit.card.debt* and *Gender* from *lm.full*. Display the summary.

A) Comment on the removal of the predicting variables by comparing *lm.red* to the full model (*lm.full*). Note any changes to the significance of the coefficients.

B) Perform a partial F-test on the new model (lm.full2) vs the previous model (lm.full), using $\alpha = 0.05$. Do you reject or fail to reject the null hypothesis. Explain your answer using the output.

C) Do the variables *credit.card.debt* and *Gender* provide predictive power? (Yes or No should suffice in conjunction w/ 11B)

```
# Code for reduced model...
lm.red<-lm(Car.Purchase.Amount~.-Gender-Credit.Card.Debt, data=buyers)
summary(lm.full)
```

```
##
## Call:
## lm(formula = Car.Purchase.Amount ~ ., data = buyers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -471.9 -202.1   17.2  205.4  463.6
##
## Coefficients:
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       -4.213e+04  1.076e+02 -391.696   <2e-16 ***
## CountryMEX         7.206e+01  4.495e+01    1.603   0.1095
## CountryUSA         5.085e+01  2.807e+01    1.812   0.0706 .
## Gender1            2.008e+01  2.184e+01    0.919   0.3583
## Age                8.397e+02  1.440e+00  583.344   <2e-16 ***
## Annual.Salary      5.623e-01  1.005e-03  559.649   <2e-16 ***
## Credit.Card.Debt   5.722e-03  3.106e-03    1.842   0.0660 .
## Net.Worth          2.893e-02  6.429e-05  449.993   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 241.1 on 492 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 1.422e+05 on 7 and 492 DF,  p-value: < 2.2e-16
```

```r
summary(lm.red)
```

```
##
## Call:
## lm(formula = Car.Purchase.Amount ~ . - Gender - Credit.Card.Debt,
##     data = buyers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -463.93 -213.40   22.53  205.84  438.09
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    -4.206e+04  1.027e+02 -409.553   <2e-16 ***
## CountryMEX      7.140e+01  4.500e+01    1.587   0.1132
## CountryUSA      5.377e+01  2.788e+01    1.928   0.0544 .
## Age             8.397e+02  1.436e+00  584.558   <2e-16 ***
## Annual.Salary   5.623e-01  1.005e-03  559.588   <2e-16 ***
## Net.Worth       2.892e-02  6.433e-05  449.596   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 241.7 on 494 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
## F-statistic: 1.981e+05 on 5 and 494 DF,  p-value: < 2.2e-16
```

```r
# Code for Partial F-Test...
anova(lm.red, lm.full)
```

```
## Analysis of Variance Table
##
## Model 1: Car.Purchase.Amount ~ (Country + Gender + Age + Annual.Salary +
##     Credit.Card.Debt + Net.Worth) - Gender - Credit.Card.Debt
## Model 2: Car.Purchase.Amount ~ Country + Gender + Age + Annual.Salary +
##     Credit.Card.Debt + Net.Worth
##   Res.Df      RSS Df Sum of Sq     F Pr(>F)
## 1    494 28863353
## 2    492 28610862  2    252491 2.171 0.1152
```

Response to Question 11A: There is no significant change in coefficient values, sign directions or statistical significance.

Response to Question 11B: Because the pvalue(0.1152) > alpha (0.05), we fail to reject the null hypothesis that the added variable coefficients are equal to 0.

Response to Question 11C: No, the added variables do NOT add predictive power, given that the other predictors are in use.

Question 12 - Predictions 4 - 3pts

Using ***lm.full*** model, what is the predicted *Car.Purchase.Amount* and corresponding ***90% prediction interval*** for a vehicle purchased in **Mexico** by a **45** year old **female** (0) with an annual salary of **$65,000**, Credit.Card.Debt of **$2,000**, and Net.Worth of **$500,000**? Provide an interpretation of your results.

Note: (Ensure you are using lm.full not lm.red.The data point has been provide.)

```r
# new observation...
newpt<-data.frame(Country='MEX', Age=45, Gender='0', Annual.Salary=65000,
                  Credit.Card.Debt=2000, Net.Worth=500000)

# Code for prediction interval...
predict(lm.full, newpt, interval="prediction", level=0.9)
```

```
##        fit      lwr      upr
## 1 46755.79 46350.64 47160.94
```

Response to Question 12:
Our model predicted the purchase amount for a car with the above characteristics to be $46,756. The 90% prediction interval is $46,351 for the lower bound and $47,161 for the upper bound. We can be 90% confident that the purchase amount for a car with these specific characteristics is between $46,351 and $47,161.

The End.