**Module Assignment**

**Module 5**

**QMB-6304 Analytical Methods for Business**

***Chad F. Lutz***



Write a simple R script to execute the following:

**Preprocessing**

1. Load into R the data in the file "6304 Module 5 Assignment Data.xlsx". This information is on 45,425 automobiles offered for sale on Craig's List in the United States. This is your master data set. The variables in the data set are:
   1. REGION: The region (defined by Craig's List) where the car was for sale.
   2. PRICE: The asking price given in the ad for the car.
   3. YEAR: The model year of the auto for sale.
   4. MAKE: The manufacturer of the car.
   5. MODEL: The particular model of the car.
   6. CONDITION: The seller-defined condition of the car.
   7. CYLINDERS: The number of cylinders of the car's engine.
   8. FUEL: The fuel type the car uses, gasoline or diesel.
   9. ODOMETER: The odometer reading (miles) on the car.
   10. PAINT COLOR: The color of the car.

**craigs.list=import("6304 Module 5 Assignment Data.xlsx")**

**colnames(craigs.list)=tolower(make.names(colnames(craigs.list)))**

1. Take a random sample of n=90 randomly selected cars. As always, use the numerical portion of your U number for the random number seed and apply the method we've demonstrated in class. This will be your primary data set and will have following characteristics:
   1. Only vehicles with MAKE of “cadillac”.
   2. Only cars from the 2006 through 2011 model years (YEAR) inclusive.
   3. Only cars with CONDITION listed as “excellent” or “good”.
   4. Only cars with engines of 6 or 8 CYLINDERS.
   5. No cars with the PAINT COLOR of “black” or “custom”.

**cars.subset=subset(craigs.list, make == "cadillac" &**

**year >= 2006 & year <= 2011 &**

**condition %in% c("excellent","good") &**

**cylinders %in% c("6","8") &**

**paint.color != "black" &**

**paint.color != "custom")**

**set.seed(13)**

**my.cars=cars.subset[sample(1:nrow(cars.subset),90),]**

**Analysis**

1. Use the str() command to show your primary data set meets all the conditions listed Preprocessing steps 2 and 3 above.

**> str(my.cars)**

**'data.frame': 90 obs. of 10 variables:**

**$ region : chr "grand rapids, MI" "utica-rome-oneida" "western slope" "youngstown, OH" ...**

**$ price : num 13200 9500 19294 13995 10900 ...**

**$ year : chr "2007" "2007" "2008" "2010" ...**

**$ make : chr "cadillac" "cadillac" "cadillac" "cadillac" ...**

**$ model : chr "escalade" "dts" "escalade" "srx" ...**

**$ condition : chr "good" "excellent" "excellent" "excellent" ...**

**$ cylinders : num 8 8 8 6 6 6 6 8 8 6 ...**

**$ fuel : chr "gas" "gas" "gas" "gas" ...**

**$ odometer : num 107390 53000 103403 73155 99176 ...**

**$ paint.color: chr "white" "white" "white" "red" ...**

1. Use R to conduct a multiple linear regression on your random sample with PRICE as the dependent variable and ODOMETER, YEAR, CONDITION, PAINT COLOR and CYLINDERS as the independent variables. Make appropriate decisions as to whether any of the independent variables should be treated as factor variables.

**my.factors=c("cylinders","year","condition",**

**"paint.color")**

**my.cars[my.factors]=lapply(my.cars[my.factors],factor)**

**cars.out=lm(price~year+condition+cylinders+odometer+**

**paint.color,data=my.cars)**

**summary(cars.out)**

1. Report the beta coefficients and associated p values and beta coefficient confidence intervals from your model. Provide appropriate interpretations of your beta coefficients.

***The intercept is $5277, or could be interpreted as the expected value for a 2006, excellent condition, V6, blue Cadillac. The model year is expected to increase the value from the 2006 value as shown in the estimates/coefficients. Interestingly, 2008 may not increase in value over 2007 with the confidence interval crossing zero, but the rest of the model years increase value. Cars listed as excellent over good are not necessarily expected to increase value either, as the p-value is high and the confidence interval again crosses zero. Odometer and paint color have little affect on value with high p-values and confidence intervals crossing zero. Cylinders is significant with V8 expected to value $4582 more than a V6 with 95% confidence to increase value between $2661 and $6502***

**> summary(cars.out)**

**Call:**

**lm(formula = price ~ year + condition + cylinders + odometer +**

**paint.color, data = my.cars)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-7768.6 -1557.3 -98.4 2290.3 9008.5**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 5.277e+03 2.300e+03 2.294 0.02458 \***

**year2007 3.482e+03 1.452e+03 2.399 0.01894 \***

**year2008 3.152e+03 1.593e+03 1.979 0.05145 .**

**year2009 5.459e+03 1.738e+03 3.141 0.00241 \*\***

**year2010 6.429e+03 1.479e+03 4.346 4.30e-05 \*\*\***

**year2011 7.598e+03 1.676e+03 4.532 2.17e-05 \*\*\***

**conditiongood -5.759e+02 8.435e+02 -0.683 0.49688**

**cylinders8 4.582e+03 9.642e+02 4.752 9.50e-06 \*\*\***

**odometer -1.349e-02 1.098e-02 -1.229 0.22297**

**paint.colorbrown -1.225e+02 2.085e+03 -0.059 0.95332**

**paint.colorgreen -5.623e+03 3.944e+03 -1.426 0.15805**

**paint.colorgrey -2.120e+03 1.931e+03 -1.097 0.27594**

**paint.colorred 5.870e+02 1.460e+03 0.402 0.68881**

**paint.colorsilver -7.164e+02 1.461e+03 -0.490 0.62524**

**paint.colorwhite 1.454e+03 1.499e+03 0.970 0.33512**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 3487 on 75 degrees of freedom**

**Multiple R-squared: 0.5014, Adjusted R-squared: 0.4083**

**F-statistic: 5.387 on 14 and 75 DF, p-value: 5.679e-07**

**> confint(cars.out)**

**2.5 % 97.5 %**

**(Intercept) 6.948343e+02 9.859306e+03**

**year2007 5.901773e+02 6.374174e+03**

**year2008 -2.031357e+01 6.324770e+03**

**year2009 1.996391e+03 8.921485e+03**

**year2010 3.482304e+03 9.376433e+03**

**year2011 4.258403e+03 1.093753e+04**

**conditiongood -2.256336e+03 1.104501e+03**

**cylinders8 2.660857e+03 6.502447e+03**

**odometer -3.536006e-02 8.379074e-03**

**paint.colorbrown -4.276050e+03 4.031143e+03**

**paint.colorgreen -1.347923e+04 2.232904e+03**

**paint.colorgrey -5.967293e+03 1.727873e+03**

**paint.colorred -2.321756e+03 3.495805e+03**

**paint.colorsilver -3.626155e+03 2.193380e+03**

**paint.colorwhite -1.532171e+03 4.440848e+03**

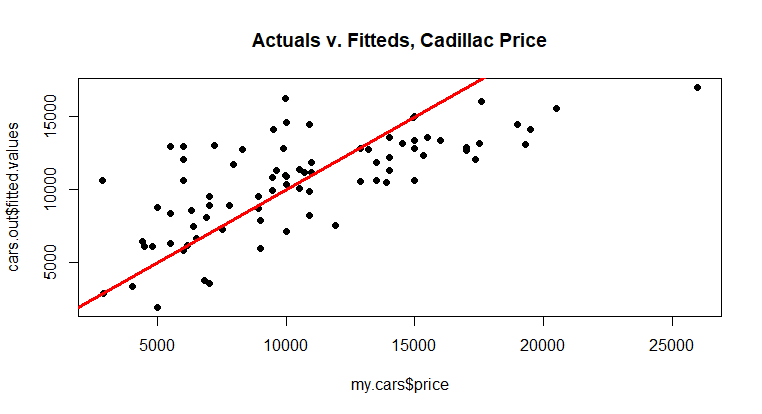
1. Conduct appropriate analyses and give interpretations to determine if your model is a good fit to the data in your primary data set.

***I don’t feel the model is a good fit due to multiple R-squared of just 0.50 and residual error of 3487. The actual v fitted plot shows a very loose fit as well. This is all using asking price anyhow, so not a very good measure.***

**plot(my.cars$price,cars.out$fitted.values,**

**pch=19,main="Actuals v. Fitteds, Cadillac Price")**

**abline(0,1,col="red",lwd=3)**



1. Assess your model’s conformance with the LINE assumptions of regression.

***The linearity is good, normality is great, EoV seems to confirm greater variance with higher priced cars, otherwise fine.***

**par(mfrow=c(2,2))**

**plot(my.cars$price,cars.out$fitted.values,**

**pch=19,main="Actuals v. Fitteds, Cadillac Price")**

**abline(0,1,col="red",lwd=3)**

**qqnorm(cars.out$residuals,pch=19,**

**main="Normality Plot, Cadillac Price")**

**qqline(cars.out$residuals,lwd=3,col="red")**

**hist(cars.out$residuals,col="red",**

**main="Residuals, Cadillac Price",**

**probability=TRUE)**

**curve(dnorm(x,mean(cars.out$residuals),**

**sd(cars.out$residuals)),**

**from=min(cars.out$residuals),**

**to=max(cars.out$residuals),**

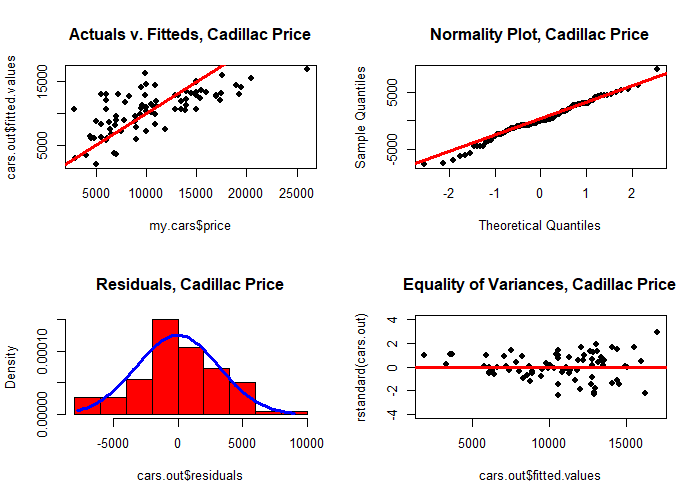
**lwd=3,col="blue",add=TRUE)**

**plot(cars.out$fitted.values,rstandard(cars.out),**

**pch=19,ylim=c(-4,4),main="Equality of Variances, Cadillac Price")**

**abline(0,0,lwd=3,col="red")**

**par(mfrow=c(1,1))**



1. Throckmorton P. Gildersleeve of Summerfield, Vermont would like to sell his red 2011 Cadillac DTS pictured above. He says the vehicle is in "excellent" condition and has 215,354 miles on the odometer. Mr. Gildersleeve has not shared details of his Cadillac’s engine with you because he thinks you know that all 2011 DTS cars had the same famous engine. If you know nothing about Cadillac engines the point is easily researched online. Use R and your model to determine what price he should ask for the car. Do you believe your pricing advice to the Great Gildersleeve is accurate and usable? Give reasoning for your conclusion.

***The predicted/fitted value for his car is $15138 and is the answer to pricing advice, but this accuracy of the model is terrible with a lower CI of $7106 and upper CI of $23170 which is to say the model has no idea where it should really be. Again, this is directly related to the model data based on asking price rather than sales price for cars.***

**maryann=data.frame(matrix(nrow=1,ncol=5))**

**colnames(maryann)=c("condition","odometer","cylinders",**

**"year","paint.color")**

**maryann[1,1]="excellent"**

**maryann[1,2]=215354**

**maryann[1,3]=8**

**maryann[1,4]=2011**

**maryann[1,5]="red"**

**colnames(maryann)=c("condition","odometer","cylinders",**

**"year","paint.color")**

**maryann[my.factors]=lapply(maryann[my.factors],factor)**

**predict(cars.out,maryann,interval = "predict")**

**> predict(cars.out,maryann,interval = "predict")**

**fit lwr upr**

**1 15138.48 7106.401 23170.57**

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above instructions, 2) the results of those instructions, and 3) any interpretations required. The first line of your script file should be a “#” comment line showing your name as it appears in Canvas. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed and submitted by the time stated on Canvas. No collaboration of any sort is allowed on this assignment.**

**#Chad Lutz**

**rm(list=ls())**

**library(rio)**

**craigs.list=import("6304 Module 5 Assignment Data.xlsx")**

**colnames(craigs.list)=tolower(make.names(colnames(craigs.list)))**

**cars.subset=subset(craigs.list, make == "cadillac" &**

**year >= 2006 & year <= 2011 &**

**condition %in% c("excellent","good") &**

**cylinders %in% c("6","8") &**

**paint.color != "black" &**

**paint.color !="custom")**

**set.seed(13)**

**my.cars=cars.subset[sample(1:nrow(cars.subset),90),]**

**str(my.cars)**

**#my.cars$cylinders=as.factor(my.cars$cylinders)**

**#my.cars$year=as.factor(my.cars$year)**

**#my.cars$condition=as.factor(my.cars$condition)**

**#my.cars$paint.color=as.factor(my.cars$paint.color)**

**#OR**

**my.factors=c("cylinders","year","condition",**

**"paint.color")**

**my.cars[my.factors]=lapply(my.cars[my.factors],factor)**

**cars.out=lm(price~year+condition+cylinders+odometer+**

**paint.color,data=my.cars)**

**summary(cars.out)**

**confint(cars.out)**

**par(mfrow=c(2,2))**

**plot(my.cars$price,cars.out$fitted.values,**

**pch=19,main="Actuals v. Fitteds, Cadillac Price")**

**abline(0,1,col="red",lwd=3)**

**qqnorm(cars.out$residuals,pch=19,**

**main="Normality Plot, Cadillac Price")**

**qqline(cars.out$residuals,lwd=3,col="red")**

**hist(cars.out$residuals,col="red",**

**main="Residuals, Cadillac Price",**

**probability=TRUE)**

**curve(dnorm(x,mean(cars.out$residuals),**

**sd(cars.out$residuals)),**

**from=min(cars.out$residuals),**

**to=max(cars.out$residuals),**

**lwd=3,col="blue",add=TRUE)**

**plot(cars.out$fitted.values,rstandard(cars.out),**

**pch=19,ylim=c(-4,4),main="Equality of Variances, Cadillac Price")**

**abline(0,0,lwd=3,col="red")**

**par(mfrow=c(1,1))**

**maryann=data.frame(matrix(nrow=1,ncol=5))**

**colnames(maryann)=c("condition","odometer","cylinders",**

**"year","paint.color")**

**maryann[1,1]="excellent"**

**maryann[1,2]=215354**

**maryann[1,3]=8**

**maryann[1,4]=2011**

**maryann[1,5]="red"**

**maryann[my.factors]=lapply(maryann[my.factors],factor)**

**predict(cars.out,maryann,interval = "predict")**