DS 700 Lesson 6 Homework

Abra Brisbin

11/22/2020

## Factors affecting the price of used cars

Load necessary packages here.

library(readr)  
library(ggformula)

## Warning: package 'ggformula' was built under R version 4.0.3

## Loading required package: ggplot2

## Loading required package: ggstance

## Warning: package 'ggstance' was built under R version 4.0.3

##   
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

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

###1a. Download Cars 2005.csv and load the data into R. Display the first few lines of data.

(Dataset: “Car Data,” submitted by Shonda Kuiper, Grinnell College. Dataset obtained from the Journal of Statistics Education (<http://www.amstat.org/publications/jse>). Accessed 3 June 2015. Used by permission of author.)

cars = read\_csv("Cars 2005.csv")

## Parsed with column specification:  
## cols(  
## Price = col\_double(),  
## Mileage = col\_double(),  
## Make = col\_character(),  
## Model = col\_character(),  
## Trim = col\_character(),  
## Type = col\_character(),  
## Cylinder = col\_double(),  
## Liter = col\_double(),  
## Doors = col\_double(),  
## Cruise = col\_double(),  
## Sound = col\_double(),  
## Leather = col\_double()  
## )

head(cars)

## # A tibble: 6 x 12  
## Price Mileage Make Model Trim Type Cylinder Liter Doors Cruise Sound  
## <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 17314. 8221 Buick Cent~ Seda~ Sedan 6 3.1 4 1 1  
## 2 17542. 9135 Buick Cent~ Seda~ Sedan 6 3.1 4 1 1  
## 3 16219. 13196 Buick Cent~ Seda~ Sedan 6 3.1 4 1 1  
## 4 16337. 16342 Buick Cent~ Seda~ Sedan 6 3.1 4 1 0  
## 5 16339. 19832 Buick Cent~ Seda~ Sedan 6 3.1 4 1 0  
## 6 15709. 22236 Buick Cent~ Seda~ Sedan 6 3.1 4 1 1  
## # ... with 1 more variable: Leather <dbl>

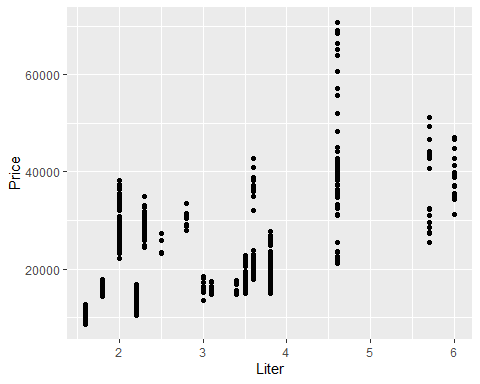
\***Note:** All of parts b-k should be done using functions from ggformula.

###1b. The variables Cruise, Sound, and Leather use 1 to represent “yes” (the car has cruise control, an upgraded sound system, or a leather interior) and 0 to represent “no”. Therefore, these variables are categorical–but R treats them as numeric, because they look like 0’s and 1’s. The function as.factor() will return a categorical variable version of its argument (the variable you put inside the parentheses). Use mutate and as.factor to convert the variables Cruise, Sound, and Leather into factor variables. (It may help to review the Reading data into R video in Lesson 5.)

cars <- cars %>%  
 mutate(Cruise = as.factor(Cruise),  
 Sound = as.factor(Sound),  
 Leather = as.factor(Leather))

###1c. Make a graph displaying the relationship between Price and Liter (a measure of engine size).

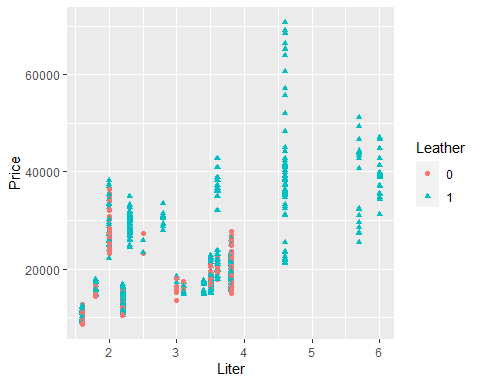
gf\_point(Price ~ Liter, data = cars)



* **Write a sentence** explaining which variable you chose as x, the independent variable, and why.
* **Write a sentence** describing the relationship between these two variables.

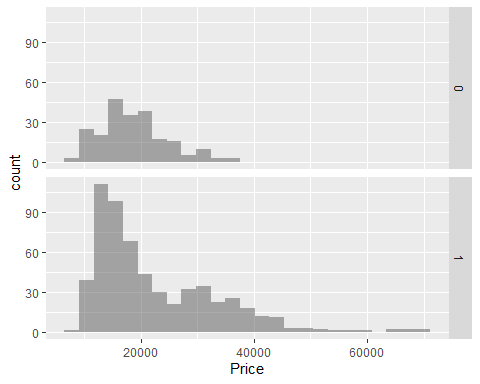
###1d. Modify your graph to include the variable Leather, in addition to Price and Liter. **Include** an informative title for your graph. There should be one set of axes.

gf\_point(Price ~ Liter, col =~ Leather, shape =~ Leather, data = cars)

 - **Write a sentence** explaining which attribute you mapped to Leather, and why.

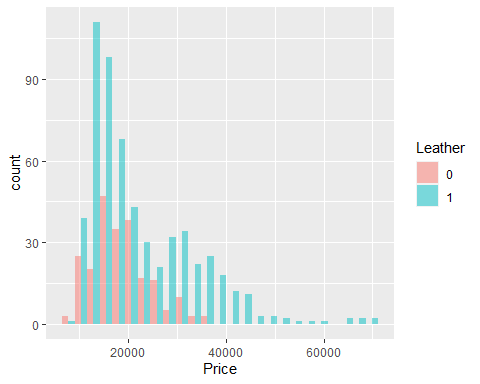
###1e. Make a graph showing the distribution of Price. Use vertically-stacked facets to display the price for cars with and without leather interiors. There should be one set of axes.

cars %>%  
 gf\_histogram( ~ Price) %>%  
 gf\_facet\_grid(Leather ~ .)



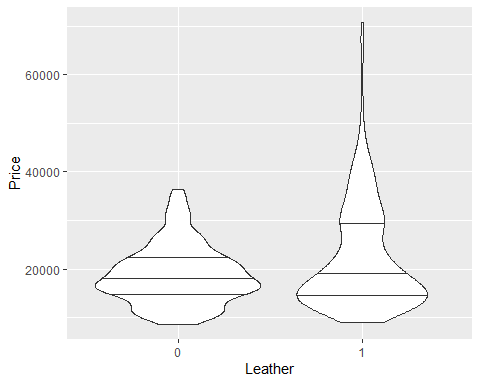
###1f. Make a graph showing the distribution of Price, with side-by-side bars for cars with and without leather. There should be one set of axes.

cars %>%  
 gf\_histogram( ~ Price, fill =~ Leather,   
 position = position\_dodge())



###1g. Make another type of graph showing the relationship between Price and Leather.

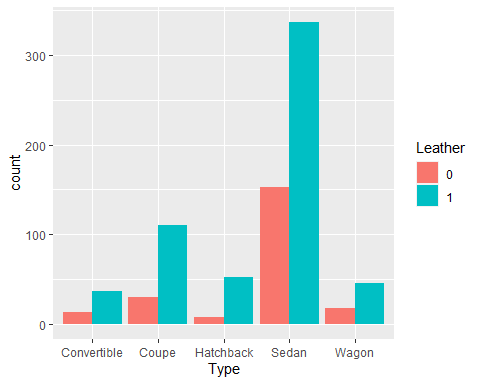
gf\_violin(Price ~ Leather, data = cars,  
 draw\_quantiles = c(.25, .5, .75))



* **Write a sentence** describing the relationship between Price and Leather.
* In your opinion, which of the graphs in parts e-g is the most effective for communicating the relationship between Price and Leather?

###1h. Make *one* graph showing the relationship between Type (Sedan, Convertible, etc.) and Leather.

gf\_bar( ~ Type, fill =~ Leather, data = cars,  
 position = position\_dodge())



* **Write 1-3 sentences** describing the relationship between Type and Leather.

## Problem 2: Regression analysis

### 2a. Use linear regression to model Price as a function of Liter.

fit = lm(Price ~ Liter, data = cars)  
fit

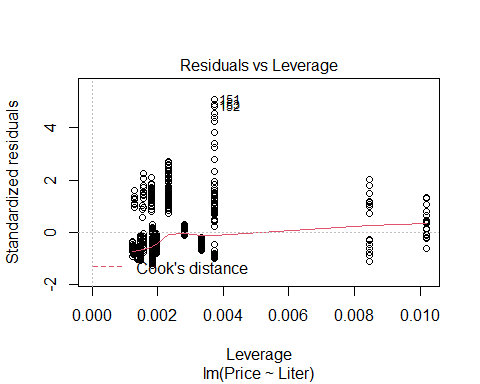
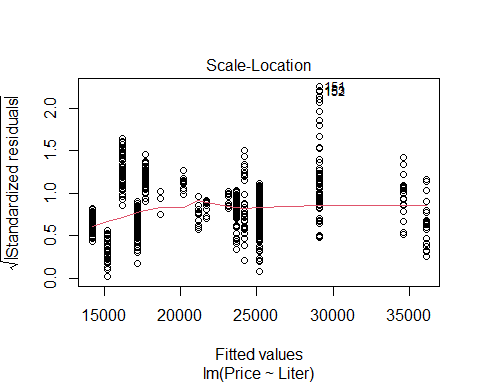
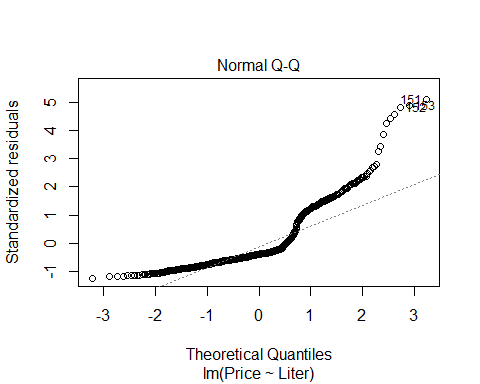
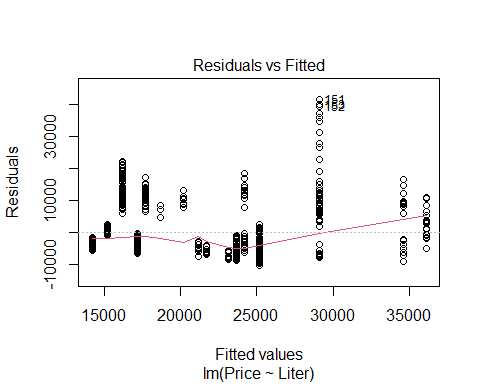
##   
## Call:  
## lm(formula = Price ~ Liter, data = cars)  
##   
## Coefficients:  
## (Intercept) Liter   
## 6186 4990

What is the equation of the line of best fit?

**Write a sentence** interpreting the slope of the line of best fit.

### 2b. Graph the residual plots of the linear model.

plot(fit)

 In 2-4 sentences, comment on the appropriateness of the linear model.

### 2c. Use linear regression to model log(Price) as a function of Liter.

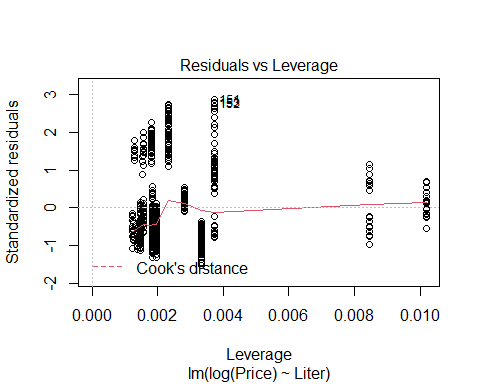
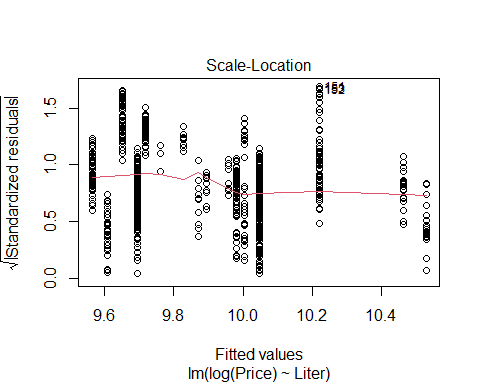
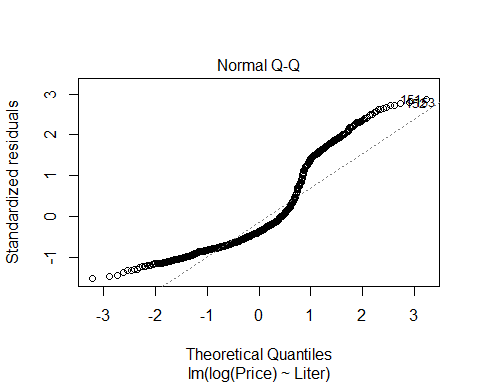
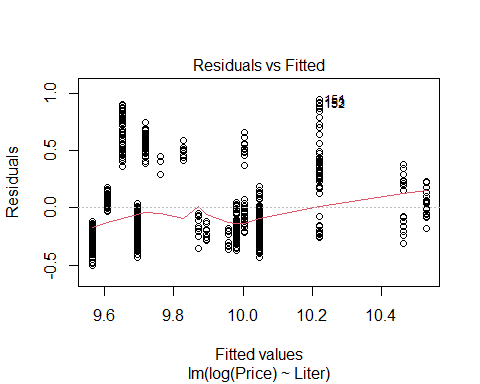
fit\_log = lm(log(Price) ~ Liter, data = cars)  
fit\_log

##   
## Call:  
## lm(formula = log(Price) ~ Liter, data = cars)  
##   
## Coefficients:  
## (Intercept) Liter   
## 9.214 0.219

Write an equation for estimated Price (not log(Price)). $log(Price) = 9.214 +0.219Liter $ $Price = e^{9.214 +0.219Liter} $ $Price = e^{9.214} e^{0.219Liter} $ $Price = e^{9.214} ^{Liter} $

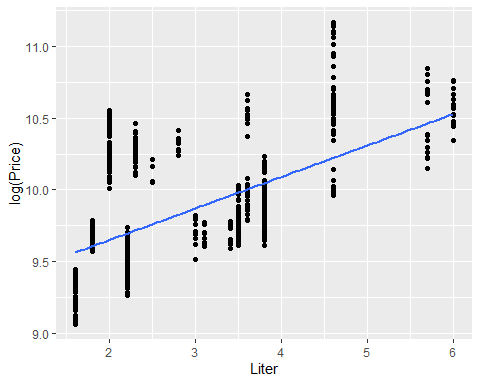
### 2d. Graph the residual plots of the linear model for log(Price). In 2-4 sentences, comment on the appropriateness of the model.

plot(fit\_log)



### 2e. Make a scatterplot of log(Price) as a function of Liter. Plot the line of best fit on the graph.

cars %>%  
 gf\_point(log(Price) ~ Liter) %>%  
 gf\_smooth(log(Price) ~ Liter, method = "lm")



### 2f. Use linear regression to model log(Price) as a function of Liter + Leather.

fit\_2var = lm(log(Price) ~ Liter + Leather, data = cars)  
fit\_2var

##   
## Call:  
## lm(formula = log(Price) ~ Liter + Leather, data = cars)  
##   
## Coefficients:  
## (Intercept) Liter Leather1   
## 9.16912 0.21645 0.07252

Write an equation for predicted log(Price) for cars without leather.

Write an equation for predicted log(Price) for cars with leather.

### 2g. Use logistic regression to model Cruise as a function of log(Price).

fit\_logistic = glm(Cruise ~ log(Price), family = "binomial", data = cars)  
fit\_logistic

##   
## Call: glm(formula = Cruise ~ log(Price), family = "binomial", data = cars)  
##   
## Coefficients:  
## (Intercept) log(Price)   
## -48.033 5.058   
##   
## Degrees of Freedom: 803 Total (i.e. Null); 802 Residual  
## Null Deviance: 899.8   
## Residual Deviance: 635.1 AIC: 639.1

Write an equation for predicted probability that a car will have cruise control. Simplify your equation using the rules , and .

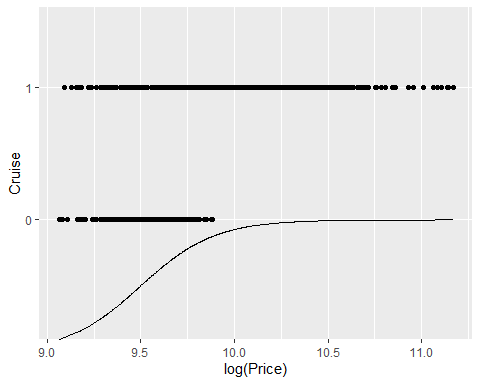
$Predicted P(y = 1) = $ $Predicted P(y = 1) = $

### 2h. Create a new data frame, cars2, containing the data from cars, plus a column containing the predicted probability that each car will have cruise control.

cars2 <- cars %>%  
 mutate(predictions = predict(fit\_logistic, type = "response"))

### 2i. Make a graph of the predicted probability of Cruise as a function of log(Price).

cars2 %>%  
 gf\_point(Cruise ~ log(Price)) %>%  
 gf\_line(predictions ~ log(Price))

 Using the graph, estimate the probability that a car will have cruise control if its log(Price) is 9.5. What if its log(Price) is 10.5?