Homework 3 part 1

Mitchell Howard

11/14/2021

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

library(ggplot2)  
library(class)  
library(caret)

## Loading required package: lattice

monet <- read.csv("~/Downloads/monet.csv")

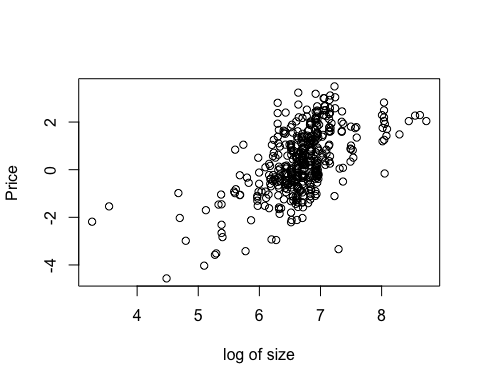
cor(monet$WIDTH, monet$PRICE)

## [1] 0.3468806

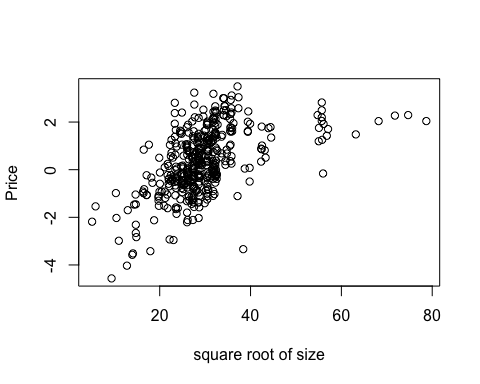
monet$SIZE <- monet$HEIGHT\*monet$WIDTH  
  
cor(monet$PRICE, monet$SIZE)

## [1] 0.3472274

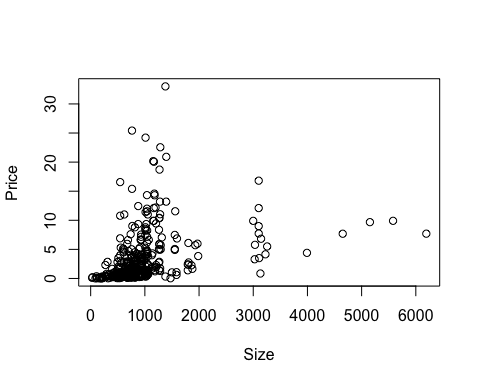
monet$logPRICE <- log(monet$PRICE)  
  
monet$logSIZE <- log(monet$SIZE)  
  
  
plot(monet$logSIZE, monet$logPRICE, xlab = "log of size", ylab = "Price")



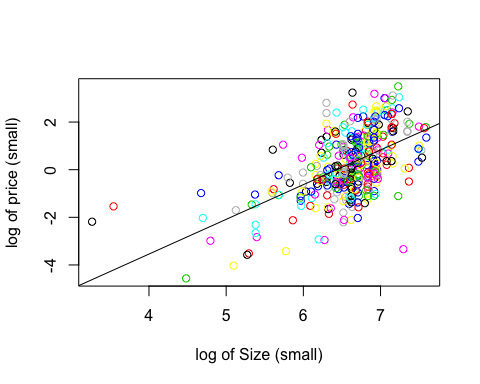
monet$sqrtSIZE <- sqrt(monet$SIZE)  
  
plot(monet$sqrtSIZE, monet$logPRICE, xlab = "square root of size", ylab = "Price")



plot(monet$SIZE, monet$PRICE, xlab = "Size", ylab = "Price")

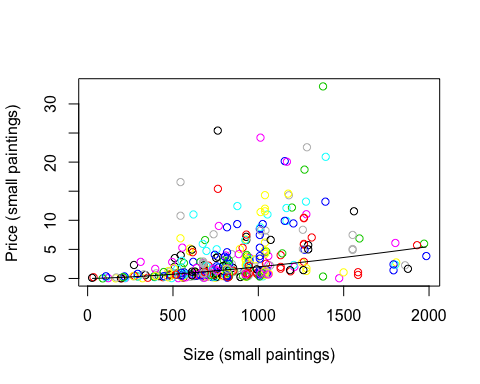


monetsmall <- monet %>% filter(SIZE < 2500)  
model <- lm(monetsmall$logPRICE~ monetsmall$logSIZE)  
  
a <- model$coefficients[1]  
b <- model$coefficients[2]  
plot(monetsmall$logSIZE, monetsmall$logPRICE, col = as.factor(monetsmall$SIZE), xlab = "log of Size (small)", ylab = "log of price (small)") + abline(model)



## integer(0)

plot(monetsmall$SIZE, monetsmall$PRICE, col = as.factor(monetsmall$SIZE), xlab = "Size (small paintings)", ylab = "Price (small paintings)")  
curve(exp(a) \* (x \*\* b), col = "black", add = TRUE)



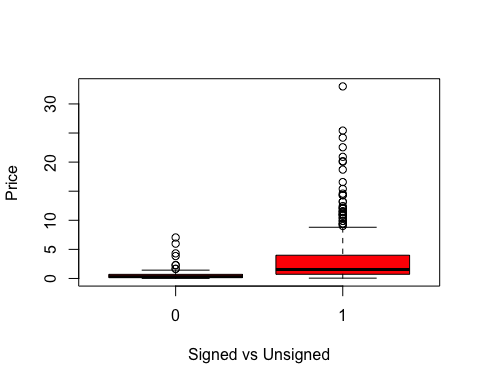
# plot(model, which = 1)   
  
  
  
summary(model)

##   
## Call:  
## lm(formula = monetsmall$logPRICE ~ monetsmall$logSIZE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.5947 -0.7198 -0.0454 0.7294 3.0006   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.3695 0.7120 -13.16 <2e-16 \*\*\*  
## monetsmall$logSIZE 1.4567 0.1073 13.58 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.111 on 411 degrees of freedom  
## Multiple R-squared: 0.3097, Adjusted R-squared: 0.308   
## F-statistic: 184.4 on 1 and 411 DF, p-value: < 2.2e-16

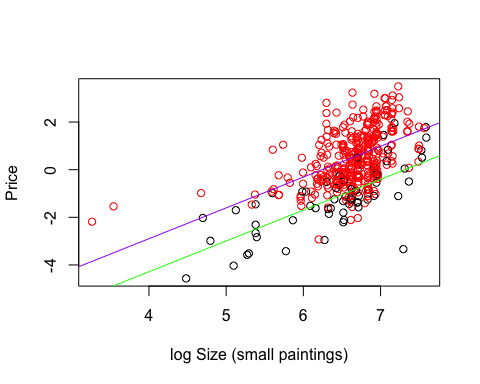
## This model is not very good, given it has an R^2 score of around .30. Typically, a well fitting model would have an R^2 score of above .90, so we have room to improve our model.

modelcomplex <- lm(monetsmall$logPRICE~ monetsmall$logSIZE + monetsmall$SIGNED)  
  
  
a1 <- modelcomplex$coefficients[1]  
b1 <- modelcomplex$coefficients[2]  
dummy <- modelcomplex$coefficients[3]

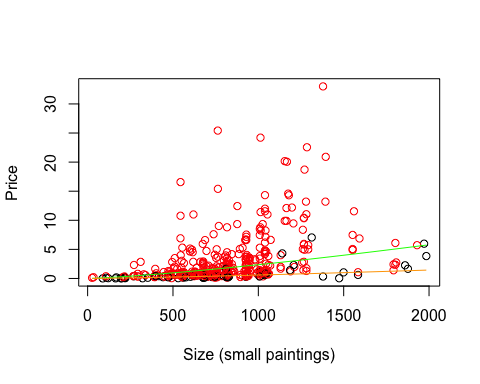
boxplot(monetsmall$PRICE ~ monetsmall$SIGNED, col = as.factor(monetsmall$SIGNED), xlab = "Signed vs Unsigned", ylab = "Price")



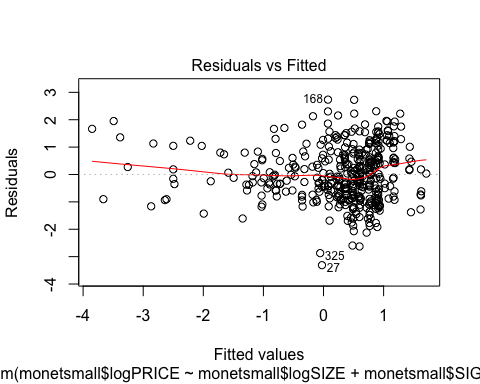
plot(monetsmall$logSIZE, monetsmall$logPRICE, col = as.factor(monetsmall$SIGNED), xlab = "log Size (small paintings)", ylab = "Price")  
abline(a1 + dummy, b1, col = "purple" )  
abline(a1, b1, col = "green")



plot(monetsmall$SIZE, monetsmall$PRICE, col = as.factor(monetsmall$SIGNED), xlab = "Size (small paintings)", ylab = "Price")  
curve(exp(a1 +dummy) \* (x \*\* b1), col = "green", add = TRUE)  
curve(exp(a1) \* (x \*\* b1), col = "orange", add = TRUE)



plot(modelcomplex, which = 1)



summary(modelcomplex)

##   
## Call:  
## lm(formula = monetsmall$logPRICE ~ monetsmall$logSIZE + monetsmall$SIGNED)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3111 -0.7461 -0.0872 0.6604 2.7333   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.45607 0.63791 -14.82 <2e-16 \*\*\*  
## monetsmall$logSIZE 1.29263 0.09746 13.26 <2e-16 \*\*\*  
## monetsmall$SIGNED 1.38753 0.13728 10.11 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9955 on 410 degrees of freedom  
## Multiple R-squared: 0.4474, Adjusted R-squared: 0.4447   
## F-statistic: 166 on 2 and 410 DF, p-value: < 2.2e-16

## Summary on Multiple regression:

The multiple regression has a better fit than the single regression with an R squared score of .4474. The best model for small painting seems to be a double log transformation. The original, non transformed data is highly skewed and follows no real pattern. By taking the log of both independent and dependent variable, we end up getting a more normally distributed set, so running linear regressions is much more straight forward.