Class Activity#5

Sunny Lee

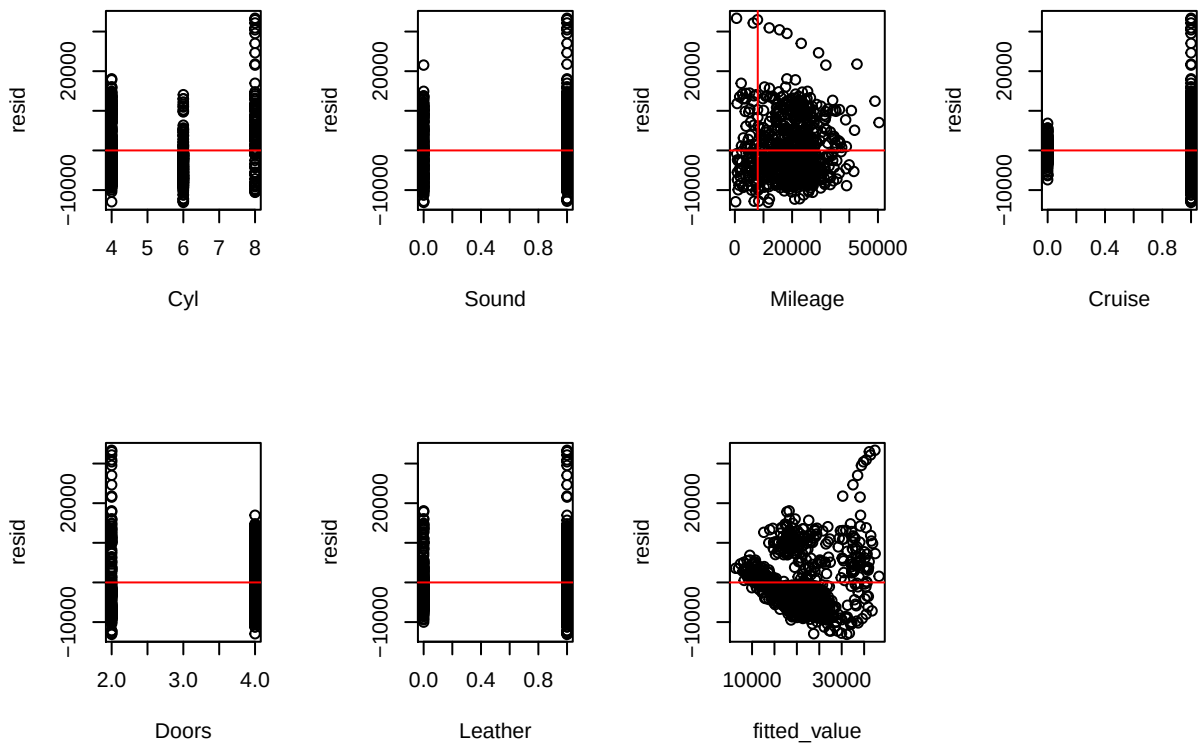
2021-02-17

1) Use the regression equation from Class activity #4, create plots of the residuals versus each explanatory variable in the model. Also create a plot of the residuals versus predicted retail price (often called a residual versus fit plot

```
cars <- read.csv("C3 Cars.csv")
attach(cars)

Best.lm <- lm(Price~Cyl+Sound+Mileage+Cruise+Doors+Leather)
resid <- Best.lm$residuals
fitted_value <- Best.lm$fitted.values

par(mfrow = c(2, 4))
plot(resid~Cyl)
abline(h = 0, col = "red")
plot(resid~Sound)
abline(h = 0, col = "red")
plot(resid~Mileage)
abline(v = 8000, col = "red")
abline(h = 0, col = "red")
plot(resid~Cruise)
abline(h = 0, col = "red")
plot(resid~Doors)
abline(h = 0, col = "red")
plot(resid~Leather)
abline(h = 0, col = "red")
plot(resid~fitted_value)
abline(h = 0, col = "red")
```



a) Does the size of the residuals tend to change as mileage changes?

Yes, if we look at the graph, we see that the residuals seem to be getting smaller as mileage increases. In other words, it seems as though the variation between the residuals and the mileage seems to be decreasing.

b) Does the size of the residuals tend to change as the predicted retail price changes? You should see patterns indicating heteroskedasticity (nonconstant variance)

Yes, in all of our plots, we see the size of the residuals seem to change. This indicates that one of our model assumptions are not met.

c) Another pattern that may not be immediately obvious from these residual plots is the right skewness seen in the residual versus mileage plot. With a pencil draw a vertical line corresponding to the mileage equal to 8000. Are the points in the residual plots balanced around the line $y = 0$?

Looking at the residuals vs mileage graph, we see that the residuals are clearly skewed below the $y = 0$ line when the mileage is equal to 8000. This clearly shows that our residuals are not normally distributed.

d) Describe any pattern seen in the other residual plots.

In the residuals vs fitted_value graph, we see the residuals increase as the fitted value increases. We also see in every single plot above that variance in the residuals do not seem to be constant.

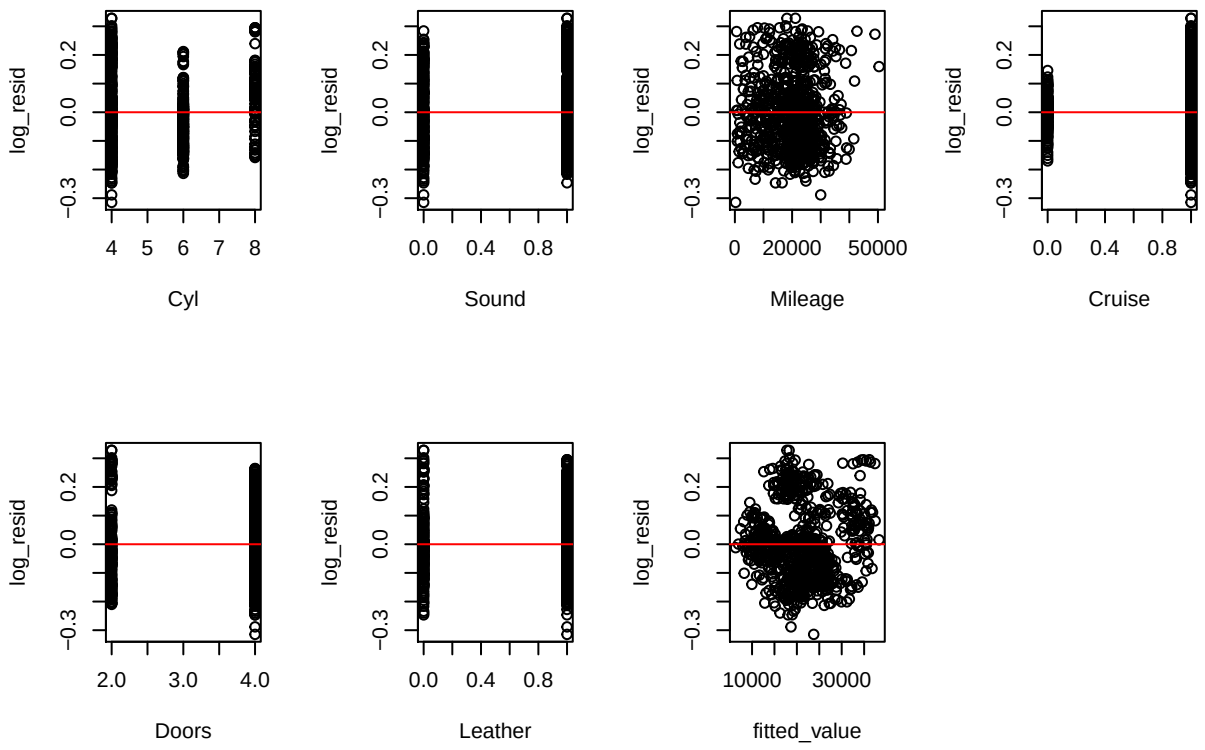
2) Transform the suggested retail price to $\log(Price)$ and $sqrt(Price)$. Create a regression models and residual plots for these transformed response variables using explanatory variables selected in class activity #4.

```
log_price <- log10(Price)
log_model <- lm(log_price~Cyl+Sound+Mileage+Cruise+Doors+Leather)
summary(log_model)
```

```
##
## Call:
## lm(formula = log_price ~ Cyl + Sound + Mileage + Cruise + Doors +
##     Leather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31460 -0.10129 -0.01396  0.07404  0.32803
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.996e+00  3.080e-02 129.744  < 2e-16 ***
## Cyl          5.654e-02  3.530e-03  16.018  < 2e-16 ***
## Sound       -3.787e-02  9.925e-03  -3.816 0.000146 ***
## Mileage     -3.206e-06  5.541e-07  -5.786 1.03e-08 ***
## Cruise       1.393e-01  1.133e-02  12.298  < 2e-16 ***
## Doors       -1.612e-02  5.361e-03  -3.007 0.002723 **
## Leather      5.273e-02  1.038e-02   5.078 4.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1285 on 797 degrees of freedom
## Multiple R-squared:  0.4836, Adjusted R-squared:  0.4797
## F-statistic: 124.4 on 6 and 797 DF,  p-value: < 2.2e-16
```

```r
log_resid <- log_model$residuals
log_predicted_price <- log_model$fitted.values

par(mfrow = c(2, 4))
plot(log_resid~Cyl)
abline(h = 0, col = "red")
plot(log_resid~Sound)
abline(h = 0, col = "red")
plot(log_resid~Mileage)
abline(h = 0, col = "red")
plot(log_resid~Cruise)
abline(h = 0, col = "red")
plot(log_resid~Doors)
abline(h = 0, col = "red")
plot(log_resid~Leather)
abline(h = 0, col = "red")
plot(log_resid~fitted_value)
abline(h = 0, col = "red")
```



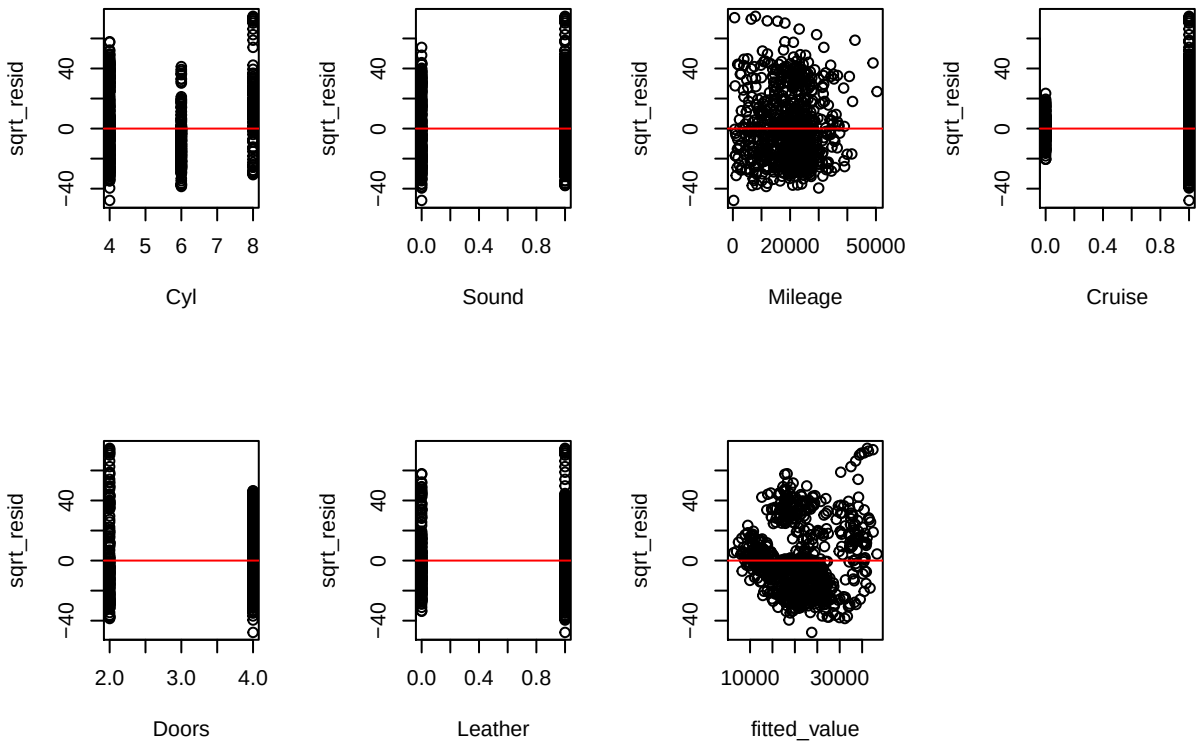

```
sqrt_price <- sqrt(Price)
sqrt_model <- lm(sqrt_price~Cyl+Sound+Mileage+Cruise+Doors+Leather)
summary(sqrt_model)
```

```
##
## Call:
## lm(formula = sqrt_price ~ Cyl + Sound + Mileage + Cruise + Doors +
##     Leather)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47.82 -17.83  -3.76  13.94  74.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.475e+01  5.435e+00  17.434  < 2e-16 ***
## Cyl          9.958e+00  6.230e-01  15.983  < 2e-16 ***
## Sound       -6.660e+00  1.752e+00  -3.802 0.000154 ***
## Mileage     -5.449e-04  9.778e-05  -5.572 3.44e-08 ***
## Cruise       2.215e+01  1.999e+00  11.080  < 2e-16 ***
## Doors       -3.679e+00  9.462e-01  -3.888 0.000109 ***
## Leather      9.934e+00  1.833e+00   5.421 7.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.67 on 797 degrees of freedom
```

```
## Multiple R-squared:  0.4689, Adjusted R-squared:  0.4649
## F-statistic: 117.3 on 6 and 797 DF,  p-value: < 2.2e-16
```

```
sqrt_resid <- sqrt_model$residuals
sqrt_predicted_price <- sqrt_model$fitted.values

par(mfrow = c(2, 4))
plot(sqrt_resid~Cyl)
abline(h = 0, col = "red")
plot(sqrt_resid~Sound)
abline(h = 0, col = "red")
plot(sqrt_resid~Mileage)
abline(h = 0, col = "red")
plot(sqrt_resid~Cruise)
abline(h = 0, col = "red")
plot(sqrt_resid~Doors)
abline(h = 0, col = "red")
plot(sqrt_resid~Leather)
abline(h = 0, col = "red")
plot(sqrt_resid~fitted_value)
abline(h = 0, col = "red")
```



a) Which transformation did the best job of reducing the heterosskedasticity and skewness in the residuals plots? Give the R^2 values of the new models.

From the above graphs, we find the log transformation did a much better job than the sqrt transformation. In each of the plots, we find the residuals seem to have very similar variation in each explanatory variable. Residuals against mileage with the log transformation also has a constant variation while the sqrt transforma-

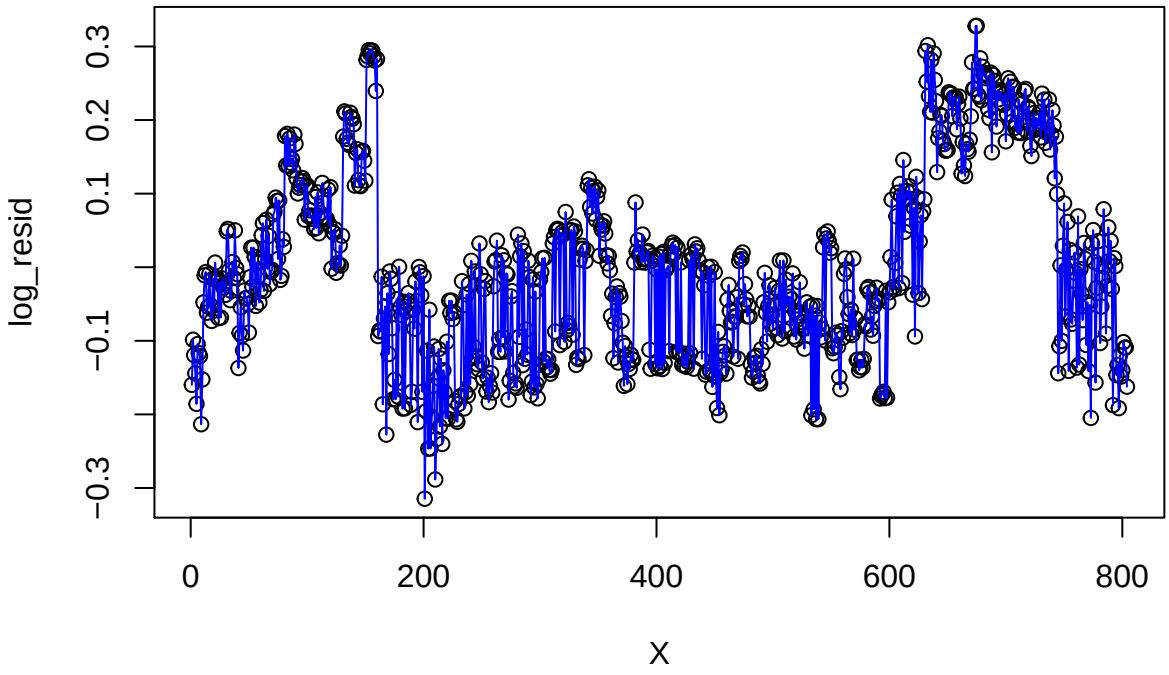
tion seesm to have decreasing variation. We got $R^2 = 0.4836$ for the log transformation, and $R^2 = 0.4689$ for the sqrt transformation, thus we can conclude that the log transformation is better for our data set than the sqrt transformation.

b) Do the best residual plots correspond to the best R^2 values? Explain.

Yes, as when we check our residual plots, we should see that none of the assumptions we made for linear regression fails. Plots with low residuals should also result in higher R^2 values as R^2 depends on the residuals.

3) Create a residual versus order plot from the price versus mileage regression line. Describe any pattern you see in the ordered residual plot.

```
# This plot is for the transformed data set
X <- seq(1, length(log_resid), 1)
plot(log_resid~X)
points(log_resid, type = "l", col = "blue")
```



Though secitons of the residual plot seem to have some kind of pattern, overall, there does not seem to be any in the residuals. Thus, we can conclude our residuals are iid.