Week 2 Assignment

David Russo

1/25/2017

* 3.7 #1
  + Intercept Null hypothesis: The average number of units sold when TV, radio, and newspaper advertising spending are 0 is equal to 0. Based on the p-value of <0.0001 we reject the null hypothesis and conclude that the average number of units sold when TV, radio, and newspaper advertising spending are 0 is different than 0.
  + TV Null hypothesis: TV advertising spending is not associated with sales of the product. Based on the p-value of <0.0001, we reject the null hypothesis and conclude that TV advertising spending is associated with sales.
  + Radio Null hypothesis: Radio advertising spending is not associated with sales of the product. Based on the p-value of <0.0001, we reject the null hypothesis and conclude that radio advertising spending is associated with sales.
  + Newspaper Null hypothesis: Newspaper advertising spending is not associated with sales of the product. Based on the p-value of 0.8599, we fail to reject the null hypothesis and conclude that newspaper advertising is likely not associated with sales.
* 3.7 #2 KNN regression and KNN Classification differ only in the form of their predicted response. KNN Classification predicts a qualitative response, while KNN regression predicts a quantitative response. In KNN classification, the predicted classification for a given test observation is the majority vote of the K nearest classifications. In KNN regression, the predicted value for a given test observation is the mean value of the K nearest responses.
* 3.7 #3
  + 1. iii is correct. This is correct because there is an interaction between gender and GPA. For low values of GPA, the +35 for females overpowers the -10*GPA. As GPA increases, the -10*GPA overpowers the +35.
    2. False. The size of a coefficient is not indicative of its significance. The significance of a coefficient depends on the size of its standard error relative to the size of the effect. In fact, with a large sample size, very small effects can be statistically significant.
* 3.7 #8

workDir <- "/Users/davidrusso/Documents/Classes/Applied Data Analysis/Assignments"  
  
auto <- read.csv(paste(workDir, "Auto.csv", sep = "/"), header = TRUE, stringsAsFactors = FALSE)  
  
auto$horsepower <- as.numeric(auto$horsepower)

## Warning: NAs introduced by coercion

+ a. lm() and summary()

lm\_hp <- lm(mpg ~ as.numeric(horsepower), data = auto)  
  
summary(lm\_hp)

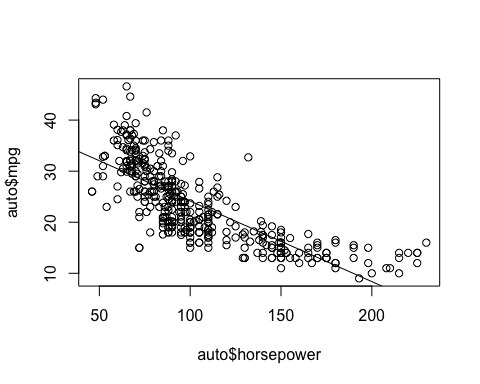
##   
## Call:  
## lm(formula = mpg ~ as.numeric(horsepower), data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5710 -3.2592 -0.3435 2.7630 16.9240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*  
## as.numeric(horsepower) -0.157845 0.006446 -24.49 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.906 on 390 degrees of freedom  
## (5 observations deleted due to missingness)  
## Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049   
## F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

+ i. There is a relationship between the predictor and the response, as evidenced by the small p-value of <2e-16.  
 + ii. The relationship between the predictor and the response is moderate, as evidenced by the R^2 of 0.6059. This indicates that 60.59% of the variation in mpg is explained by the variation in horsepower.   
 + iii. The relationship between mpg and horsepower is negative, as evidenced by the negative coefficient for horsepower.   
 + iv.

new\_data <- data.frame(horsepower = 98)  
pred\_val <- predict(lm\_hp, newdata = new\_data)  
pred\_val\_ci <- predict(lm\_hp, newdata = new\_data, interval = "confidence")  
pred\_val\_pi <- predict(lm\_hp, newdata = new\_data, interval = "prediction")

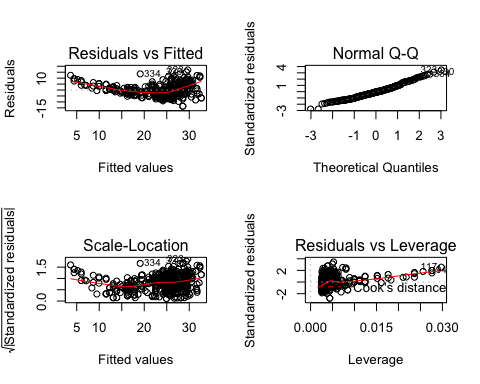
+ 1. The predicted mpg at a horsepower of 98 is 24.48 mpg  
 + 2. The confidence interval for this prediction is [23.97, 24.96]  
 + 3. The prediction interval for this prediction is [14.81, 34.12]  
   
+ b. plot()

plot(auto$horsepower, auto$mpg)  
abline(lm\_hp$coefficients[1], lm\_hp$coefficients[2])



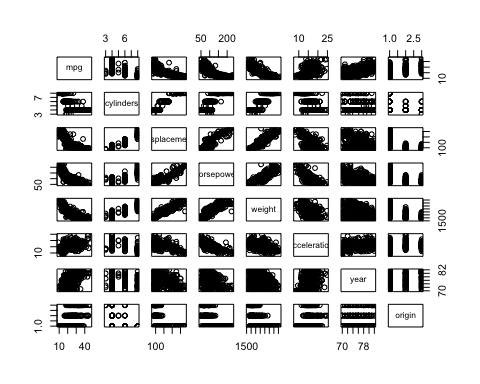
+ c. plot() for diagnostics

par(mfrow = c(2,2))  
plot(lm\_hp)

 + The residual plot indicates that the variance is non-constant. As the fitted values get larger, the residuals get larger. This is a violation of the assumption of constant variance. This also provides evidence that the relationship between horsepower and mpg is non-linear. The qq-plot indicates some lack of normality at the tails of the distribution.

* 3.7 #9
  + 1. pairs()

pairs(auto[, !names(auto) %in% "name"])



+ b. cor()

cor(auto[!is.na(auto$horsepower), !names(auto) %in% "name"])

## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  
## cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273  
## displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  
## horsepower -0.7784268 0.8429834 0.8972570 1.0000000 0.8645377  
## weight -0.8322442 0.8975273 0.9329944 0.8645377 1.0000000  
## acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392  
## year 0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199  
## origin 0.5652088 -0.5689316 -0.6145351 -0.4551715 -0.5850054  
## acceleration year origin  
## mpg 0.4233285 0.5805410 0.5652088  
## cylinders -0.5046834 -0.3456474 -0.5689316  
## displacement -0.5438005 -0.3698552 -0.6145351  
## horsepower -0.6891955 -0.4163615 -0.4551715  
## weight -0.4168392 -0.3091199 -0.5850054  
## acceleration 1.0000000 0.2903161 0.2127458  
## year 0.2903161 1.0000000 0.1815277  
## origin 0.2127458 0.1815277 1.0000000

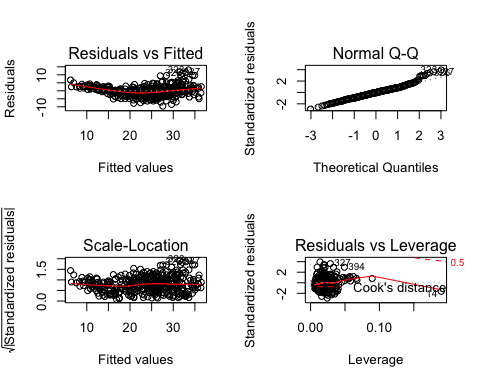
+ c. lm()

lm\_all <- lm(mpg ~ ., data = auto[, !names(auto) %in% "name"])  
summary(lm\_all)

##   
## Call:  
## lm(formula = mpg ~ ., data = auto[, !names(auto) %in% "name"])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5903 -2.1565 -0.1169 1.8690 13.0604   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.218435 4.644294 -3.707 0.00024 \*\*\*  
## cylinders -0.493376 0.323282 -1.526 0.12780   
## displacement 0.019896 0.007515 2.647 0.00844 \*\*   
## horsepower -0.016951 0.013787 -1.230 0.21963   
## weight -0.006474 0.000652 -9.929 < 2e-16 \*\*\*  
## acceleration 0.080576 0.098845 0.815 0.41548   
## year 0.750773 0.050973 14.729 < 2e-16 \*\*\*  
## origin 1.426141 0.278136 5.127 4.67e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.328 on 384 degrees of freedom  
## (5 observations deleted due to missingness)  
## Multiple R-squared: 0.8215, Adjusted R-squared: 0.8182   
## F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16

+ i. There is an overall relationship between the predictors and mpg, as evidenced by the p-value for the F-test in the summary. The p-value of <2.2e-16 indicates that at least one of the coefficients is different from zero.  
 + ii. The following predictors appear to be related to mpg: displacement, weight, year, and origin. Each of these p-values is less than 0.05.   
 + iii. The year coefficient of 0.75 indicates that for each additional year (i.e., the newer the car is), mileage increases by 0.75 miles per gallon.   
  
+ d. plot()

par(mfrow = c(2,2))  
plot(lm\_all)



+ The residual plot indicates that a linear regression may not be the most appropriate choice in model as the relationship between the predictors and mpg may not be linear. Furthermore, there is evidence that the assumption of constant variance is violated. As the fitted values get larger, so do the residuals. The residual plot indicates that there may be three outliers towards the higher end of the fitted values corresponding to data points 323, 326, and 327. The leverage plot indicates that points 14, 327, and 394 may have unusually high leverage.   
  
+ e. interactions

lm\_int1 <- lm(mpg ~ displacement + weight + year + origin + year\*displacement, data = auto)  
summary(lm\_int1)

##   
## Call:  
## lm(formula = mpg ~ displacement + weight + year + origin + year \*   
## displacement, data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.9992 -1.9574 0.0096 1.5654 13.0124   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.779e+01 7.210e+00 -8.016 1.28e-14 \*\*\*  
## displacement 2.260e-01 3.487e-02 6.481 2.74e-10 \*\*\*  
## weight -6.156e-03 5.354e-04 -11.497 < 2e-16 \*\*\*  
## year 1.295e+00 9.415e-02 13.753 < 2e-16 \*\*\*  
## origin 9.770e-01 2.561e-01 3.814 0.000159 \*\*\*  
## displacement:year -3.027e-03 4.759e-04 -6.361 5.62e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.192 on 391 degrees of freedom  
## Multiple R-squared: 0.8358, Adjusted R-squared: 0.8337   
## F-statistic: 397.9 on 5 and 391 DF, p-value: < 2.2e-16

lm\_int2 <- lm(mpg ~ displacement + weight + year + origin + year:weight, data = auto)  
summary(lm\_int2)

##   
## Call:  
## lm(formula = mpg ~ displacement + weight + year + origin + year:weight,   
## data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.9468 -1.8978 -0.0939 1.6299 12.1542   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.092e+02 1.277e+01 -8.550 2.81e-16 \*\*\*  
## displacement -1.475e-04 4.540e-03 -0.032 0.974094   
## weight 2.659e-02 4.523e-03 5.879 8.86e-09 \*\*\*  
## year 1.986e+00 1.699e-01 11.689 < 2e-16 \*\*\*  
## origin 9.057e-01 2.528e-01 3.583 0.000382 \*\*\*  
## weight:year -4.385e-04 5.926e-05 -7.400 8.41e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.14 on 391 degrees of freedom  
## Multiple R-squared: 0.841, Adjusted R-squared: 0.839   
## F-statistic: 413.7 on 5 and 391 DF, p-value: < 2.2e-16

lm\_int3 <- lm(mpg ~ displacement + weight + year\*origin, data = auto)  
summary(lm\_int3)

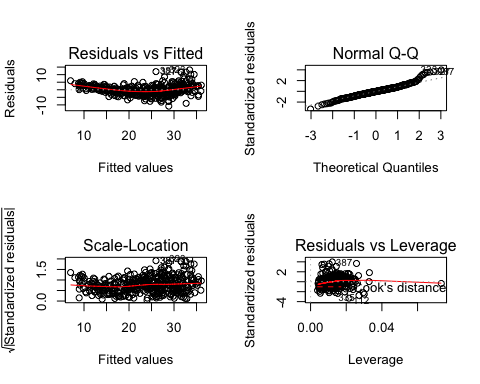
##   
## Call:  
## lm(formula = mpg ~ displacement + weight + year \* origin, data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7506 -1.8980 -0.1495 1.7274 12.3853   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.992e+00 8.784e+00 0.910 0.363499   
## displacement 1.995e-03 4.844e-03 0.412 0.680700   
## weight -6.463e-03 5.511e-04 -11.726 < 2e-16 \*\*\*  
## year 4.319e-01 1.120e-01 3.857 0.000134 \*\*\*  
## origin -1.474e+01 4.682e+00 -3.148 0.001768 \*\*   
## year:origin 2.056e-01 6.017e-02 3.417 0.000699 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.304 on 391 degrees of freedom  
## Multiple R-squared: 0.824, Adjusted R-squared: 0.8218   
## F-statistic: 366.2 on 5 and 391 DF, p-value: < 2.2e-16

+ Each of the interactions considered were significant, as evidenced by each of their p-values being significantly smaller than 0.05.   
   
+ f. Transformations

lm\_log\_disp <- lm(mpg ~ log(displacement) + weight + year + origin, data = auto)  
summary(lm\_log\_disp)

##   
## Call:  
## lm(formula = mpg ~ log(displacement) + weight + year + origin,   
## data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.8335 -1.9319 -0.0711 1.8227 13.1285   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.4093597 5.8207321 -0.929 0.35329   
## log(displacement) -2.9850847 1.0000807 -2.985 0.00302 \*\*   
## weight -0.0044972 0.0005688 -7.907 2.71e-14 \*\*\*  
## year 0.7413597 0.0479410 15.464 < 2e-16 \*\*\*  
## origin 0.7859067 0.2834637 2.773 0.00583 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.318 on 392 degrees of freedom  
## Multiple R-squared: 0.8221, Adjusted R-squared: 0.8202   
## F-statistic: 452.7 on 4 and 392 DF, p-value: < 2.2e-16

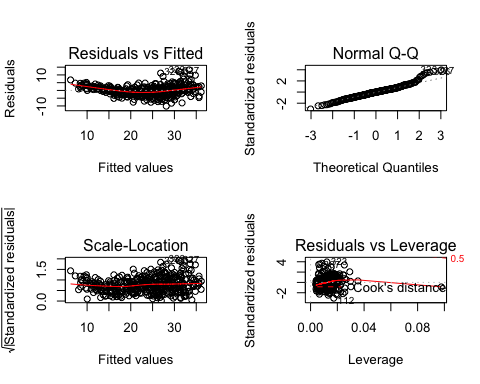
par(mfrow = c(2,2))  
plot(lm\_log\_disp)



lm\_sqrt\_disp <- lm(mpg ~ sqrt(displacement) + weight + year + origin, data = auto)  
summary(lm\_sqrt\_disp)

##   
## Call:  
## lm(formula = mpg ~ sqrt(displacement) + weight + year + origin,   
## data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.1328 -2.0770 -0.0858 1.7631 13.2268   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.717e+01 4.184e+00 -4.103 4.96e-05 \*\*\*  
## sqrt(displacement) -1.185e-01 1.472e-01 -0.805 0.42128   
## weight -5.598e-03 5.843e-04 -9.582 < 2e-16 \*\*\*  
## year 7.528e-01 4.903e-02 15.355 < 2e-16 \*\*\*  
## origin 1.076e+00 2.754e-01 3.907 0.00011 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.353 on 392 degrees of freedom  
## Multiple R-squared: 0.8183, Adjusted R-squared: 0.8165   
## F-statistic: 441.4 on 4 and 392 DF, p-value: < 2.2e-16

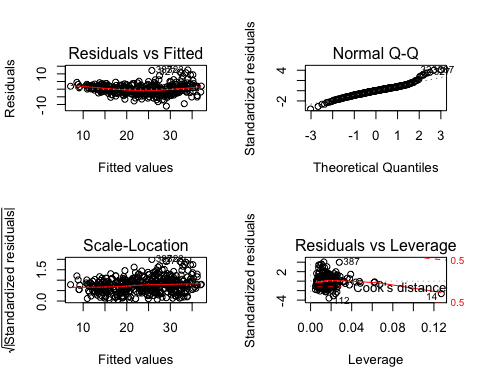
par(mfrow = c(2,2))  
plot(lm\_sqrt\_disp)



lm\_sqrd\_disp <- lm(mpg ~ displacement + I(displacement^2) + weight + year + origin, data = auto)  
summary(lm\_sqrd\_disp)

##   
## Call:  
## lm(formula = mpg ~ displacement + I(displacement^2) + weight +   
## year + origin, data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.0941 -1.8668 0.0412 1.7649 12.5834   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.536e+01 3.750e+00 -4.097 5.09e-05 \*\*\*  
## displacement -7.662e-02 1.159e-02 -6.611 1.25e-10 \*\*\*  
## I(displacement^2) 1.508e-04 1.952e-05 7.725 9.49e-14 \*\*\*  
## weight -5.391e-03 5.435e-04 -9.919 < 2e-16 \*\*\*  
## year 8.138e-01 4.638e-02 17.548 < 2e-16 \*\*\*  
## origin 3.752e-01 2.713e-01 1.383 0.167   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.123 on 391 degrees of freedom  
## Multiple R-squared: 0.8428, Adjusted R-squared: 0.8408   
## F-statistic: 419.2 on 5 and 391 DF, p-value: < 2.2e-16

par(mfrow = c(2,2))  
plot(lm\_sqrd\_disp)

 Applying the three transformations to dispalcement does not appear to have much of an effect on model fit. In each case, the residual plot does not show much evidence of improved fit. Furthermore, each plot illustrates that the assumption of constant variance is violated as the residuals increase with the fitted values. Lastly, the assumption of normality is violated with each model, as evidenced by the qq-plots. The log and square of displacement are significant, while the square root of displacement is not significant. This indicates that either a polynomial fit with displacement or using the log of displacement in other models may yield a better fit.