ASSIGNMENT 8

Brandon Lampe
STAT 527
Advanced Data Analysis I

December 2, 2014

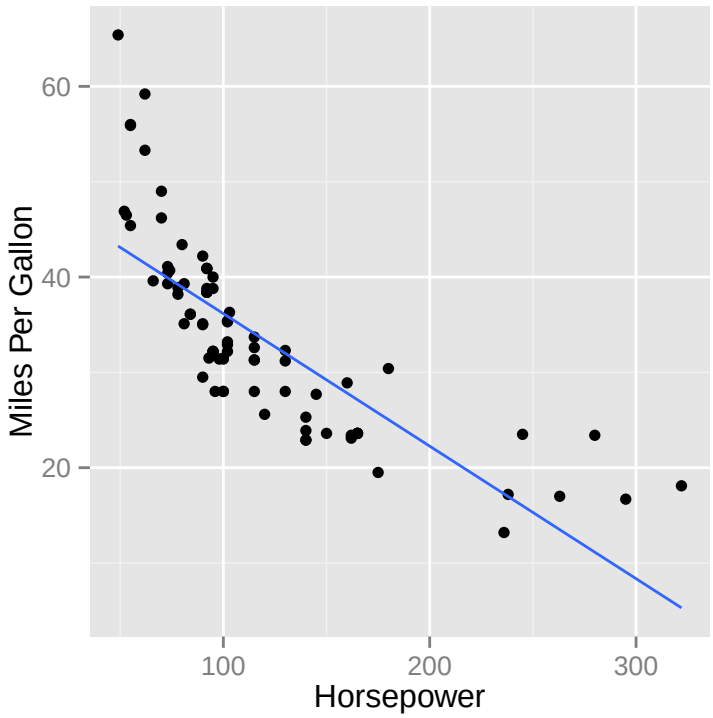
1 Gas milage and automobile horsepower:

1(a) (10 pts) Scatter Plots of Data.

```
# # read the table in as a data.frame
# # cars <- read.table("http://statacumen.com/teach/ADA1/ADA1_HW_08_F14-1.txt", header=TRUE)
#
# # write.table(cars, file = "/Users/Lampe/Documents/UNM_Courses/STAT-527_ADAI_ERHARDT/HW08/cars.txt",
# #       sep = ",", col.names = TRUE)

cars <- read.csv("/Users/Lampe/Documents/UNM_Courses/STAT-527_ADAI_ERHARDT/HW08/cars.txt", header = TRUE)
```

```
# plot mpg = f(hp)
p1a <- ggplot(cars, aes(x = hp, y = mpg))
p1a <- p1a + geom_point()
p1a <- p1a + labs(y = "Miles Per Gallon", x = "Horsepower")
p1a <- p1a + geom_smooth(method = lm, se = FALSE)
p1a
```

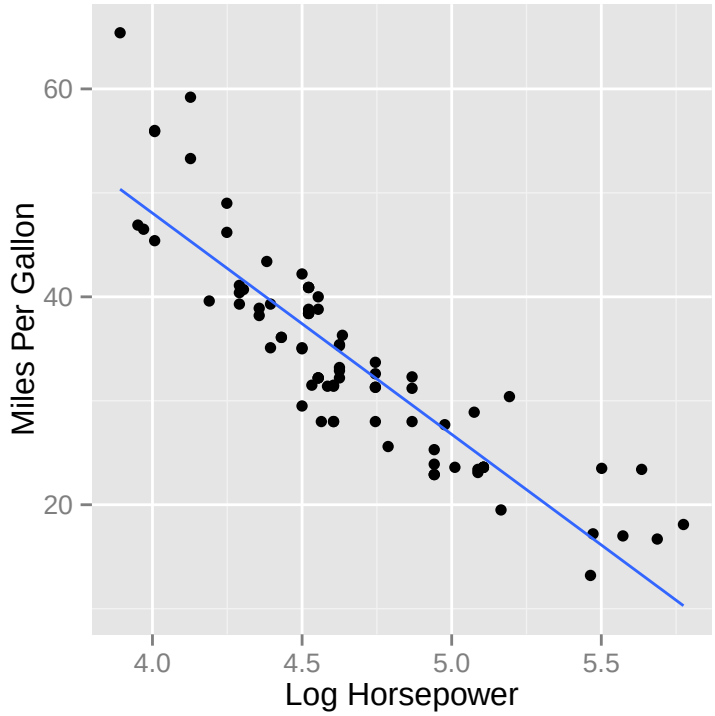



```
lm.mpg.hp <- lm(mpg ~ hp, data = cars)
sum.lm.mpg.hp <- summary(lm.mpg.hp)
summary(lm.mpg.hp)
##
```

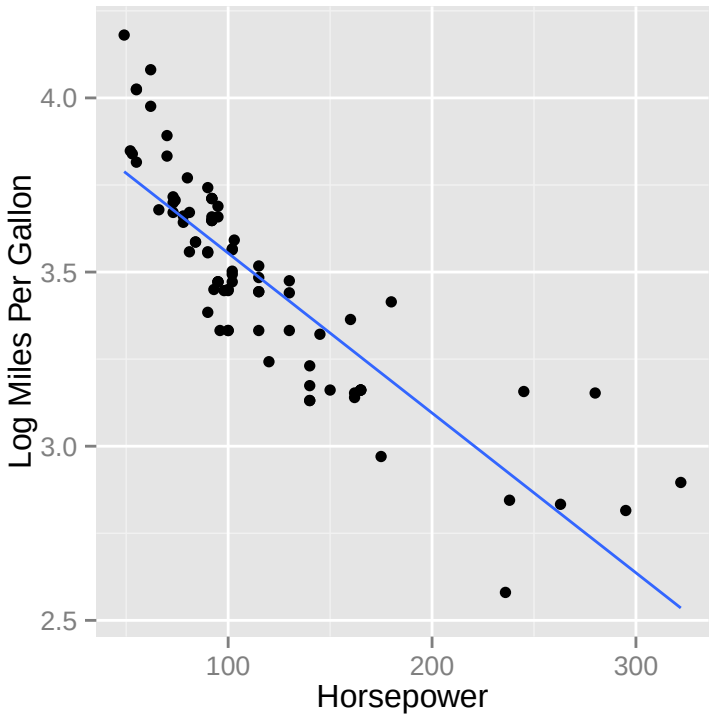
```
## Call:
## lm(formula = mpg ~ hp, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7198 -4.1224 -0.9077  3.1009 22.1461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.06608    1.56949   31.90   <2e-16 ***
## hp          -0.13902    0.01207  -11.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.174 on 80 degrees of freedom
## Multiple R-squared:  0.6239,Adjusted R-squared:  0.6192
## F-statistic: 132.7 on 1 and 80 DF,  p-value: < 2.2e-16
```

```
cars$log_hp <- log(cars$hp)
cars$log_mpg <- log(cars$mpg)
cars$log_wt <- log(cars$wt)

p1b <- ggplot(cars, aes(x = log_hp, y = mpg))
p1b <- p1b + geom_point()
p1b <- p1b + labs(y = "Miles Per Gallon", x = "Log Horsepower")
p1b <- p1b + geom_smooth(method = lm, se = FALSE)
p1b
```



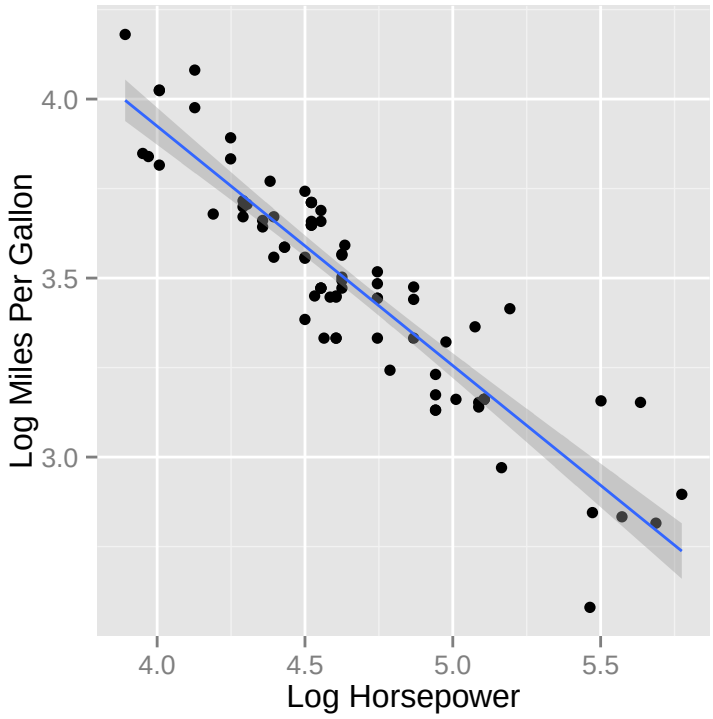

```
p1c <- ggplot(cars, aes(x = hp, y = log_mpg))
p1c <- p1c + geom_point()
p1c <- p1c + labs(y = "Log Miles Per Gallon", x = "Horsepower")
p1c <- p1c + geom_smooth(method = lm, se = FALSE)
p1c
```




```
lm.mpg.loghp <- lm(mpg ~ log_hp, data = cars)
sum.lm.mpg.loghp <- summary(lm.mpg.loghp)
summary(lm.mpg.loghp)
##
## Call:
## lm(formula = mpg ~ log_hp, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0422 -2.9050 -0.9174  2.5020 15.0468
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  133.169      5.855   22.75   <2e-16 ***
## log_hp       -21.279      1.249  -17.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.678 on 80 degrees of freedom
## Multiple R-squared:  0.784,Adjusted R-squared:  0.7813
## F-statistic: 290.5 on 1 and 80 DF,  p-value: < 2.2e-16
lm.logmpg.hp <- lm(log_mpg ~ hp, data = cars)
sum.lm.logmpg.hp <- summary(lm.logmpg.hp)
summary(lm.logmpg.hp)
##
## Call:
## lm(formula = log_mpg ~ hp, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35003 -0.10508 -0.00974  0.07242  0.42440
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0132294  0.0401238  100.02   <2e-16 ***
```

```
## hp          -0.0045889  0.0003085  -14.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1578 on 80 degrees of freedom
## Multiple R-squared:  0.7344,Adjusted R-squared:  0.7311
## F-statistic: 221.2 on 1 and 80 DF,  p-value: < 2.2e-16
```

```
p1d <- ggplot(cars, aes(x = log_hp, y = log_mpg))
p1d <- p1d + geom_point()
p1d <- p1d + labs(y = "Log Miles Per Gallon", x = "Log Horsepower")
p1d <- p1d + geom_smooth(method = lm)
p1d
```

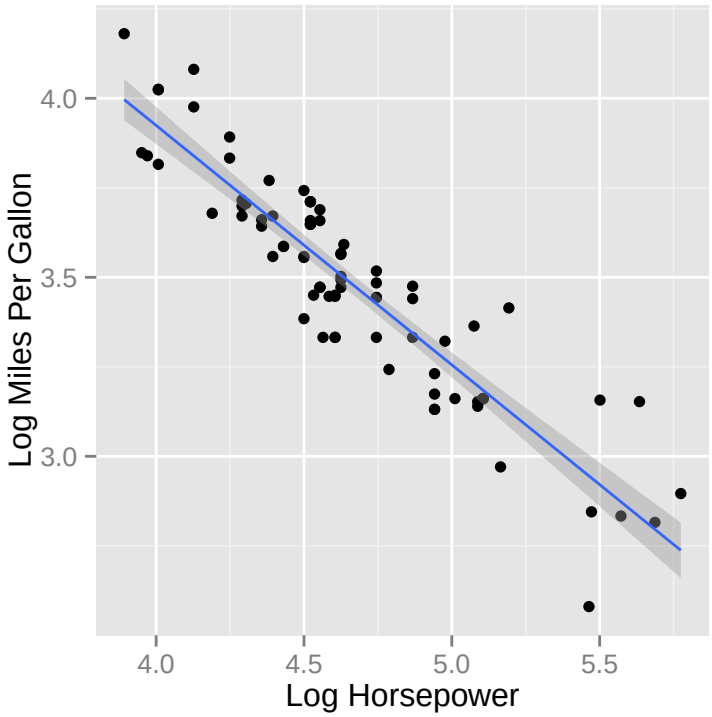



```
lm.logmpg.loghp <- lm(log_mpg ~ log_hp, data = cars)
sum.lm.logmpg.loghp <- summary(lm.logmpg.loghp)
summary(lm.logmpg.loghp)
##
## Call:
## lm(formula = log_mpg ~ log_hp, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36502 -0.08179 -0.02332  0.09198  0.32183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.59911    0.15492   42.60   <2e-16 ***
## log_hp      -0.66874    0.03304  -20.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1238 on 80 degrees of freedom
## Multiple R-squared:  0.8366,Adjusted R-squared:  0.8346
## F-statistic: 409.7 on 1 and 80 DF,  p-value: < 2.2e-16
```

Based on these graphical representations of data, a log-log transformation of the data (both x and y) is most appropriate for a straight-line regression. Additionally, the R-squared value for the log-log transformation is the highest of the four fits at 0.84, which means the log(mpg) vs log(hp) transformation most closely follows a straight-line regression model.

1(b) (10 pts) Using the more appropriate of the two pairs of variables from (a), that is original or log- transformed variables, fit a simple linear regression model.

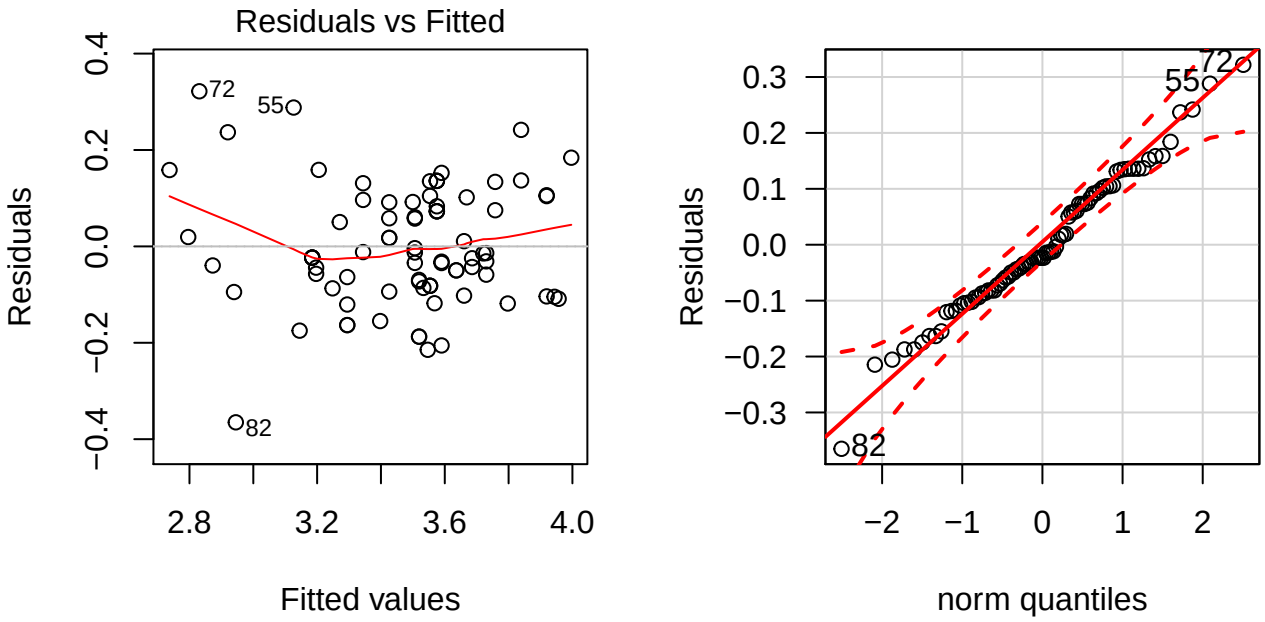
The log-log transformation was determined to be the most appropriate form for a linear regression model, results of this model and the data are shown below. A summary of this regression was provided in my response to part A of this problem.



Residuals of this linear regression model and a bootstrap sampling of the residuals are shown below. Additionally, a plot of residuals and a Q-Q plot of norm quantities are shown.

```
par(mfrow = c(1,2))
plot(lm.logmpg.loghp, which = 1)
abline(h = 0, col = "gray75")

qqPlot(lm.logmpg.loghp$residuals, las = 1, id.n = 3,
       ylab = "Residuals")
## 82 72 55
##  1 82 81
```



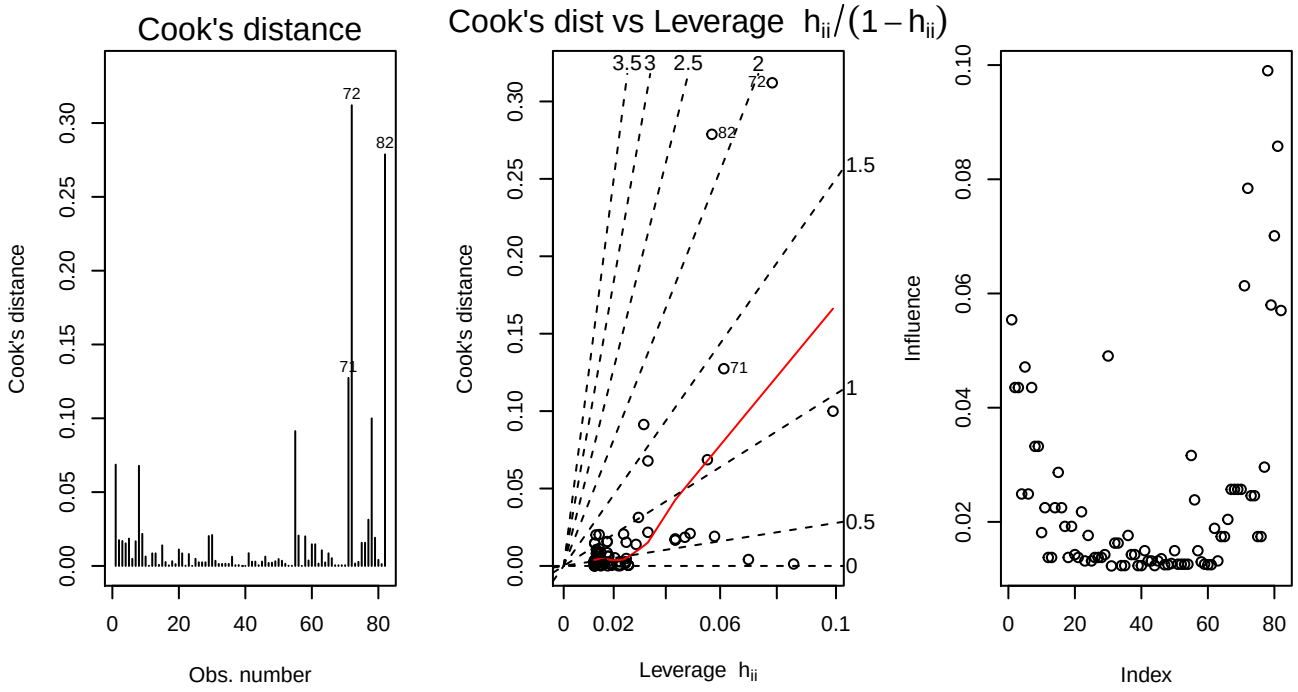
Interpretation of Residuals:

- There does not appear to be curvature associated with the residual plots. From this observation, the quality of the fit does not vary with the independent variable (log(horsepower)).
- No obvious outliers exist.
- The Q-Q plot does not highlight any skewness in the resdiuals.
- No obvious pattern for the residuals exists.

Based on these observations, the residuals appear random and normally distributed.

1(c) (5 pts) Investigate the leverages and Cooks D. Use the 3p/n cutoff for large leverages, and the cutoff of 1 for large Cooks D values.

```
par(mfrow=c(1,3))
plot(lm.logmpg.loghp,which = c(4,6))
plot(influence(lm.logmpg.loghp)$hat, ylab = "Influence")
```



Plots of Cook's distance, leverage versus Cook's distance, and influence / leverage are shown above. Based on these visual analyses, no residual has a Cook's distance greater than 1, but three residual values have large leverages based on the 3p/n cutoff (leverage of 0.74). No Values have both a Cook's distance greater than one and a leverage greater than 3p/n; therefore I do not think any observations have undue influence on the model fit.

1(d) (10 pts) Assuming the model fits well, present and interpret the ANOVA table and R2 value.

```
mpg.anova <- anova(lm.logmpg.loghp)
mpg.anova
## Analysis of Variance Table
##
## Response: log_mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## log_hp     1 6.2780  6.2780  409.68 < 2.2e-16 ***
## Residuals 80 1.2259  0.0153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
f.crit <- qf(1 - 0.05/2, mpg.anova[1,1], mpg.anova[2,1])
f.crit
## [1] 5.218354
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
log_hp	1.000	6.278	6.278	409.682	0.000
Residuals	80.000	1.226	0.015		

The ANOVA F-test may be used to determine if β_1 (slope of regression line) is equal to zero; i.e., the mileage is not related to horsepower. That is, $H_0 : \beta_1 = 0$ against the alternative $H_A : \beta_1 \neq 0$. Because the F value of 410 is much larger than the critical F value of 5.2 and because the p-value is much less than 0.05, I reject the null hypothesis in favor of the alternative. That is, the variation of mileage is related to vehicle horsepower. The Sum of Squares divided by the degrees of freedom for the residuals ($s^2_{X|Y}$) was 0.015.

The R^2 value for the linear model was 0.837, which indicates that 83.7% of the variation of mileage can be explained by variation of horsepower.

1(e) (10pts) Present the parameter estimate table and estimated regression equation. State what the hypothesis test is related to the log(hp) line in the parameter estimate table. State the conclusion of the hypothesis test. Interpret the slope coeffcient in the context of the model.

```
coef <- sum.lm.logmpg.loghp$coefficients
```

Parameter Estimate Table:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.599	0.155	42.598	0.000
log_hp	-0.669	0.033	-20.241	0.000

Estimated Regression Equation: $\log(\text{Mileage}) = -0.699 * \log(\text{Horsepower}) + 6.599$. That is, $b_0 = 6.599$ and $b_1 = -0.699$.

Hypothesis tests of the regression equation parameters β_0 (intercept) and β_1 (slope) both have a null hypothesis that these values are equal to zero. That is, $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$ with the alternative $H_A : \beta_0$ and $\beta_1 \neq 0$.

Slope of the regression line indicates that for every one unit of increase in log(Horsepower) the log(Mileage) will decrease by 0.699.

1(f) (5pts) Using the R2 statistic and the slope of the regression line, what is the correlation between log(hp) and log(mpg)?

As was done previously: the R^2 value for the linear model was 0.837, which indicates that 83.7% of the variation of mileage can be explained by variation of horsepower. The estimated regressioneEquation is: $\log(\text{Mileage}) = -0.699 * \log(\text{Horsepower}) + 6.599$. That is, $b_0 = 6.599$ and $b_1 = -0.699$.

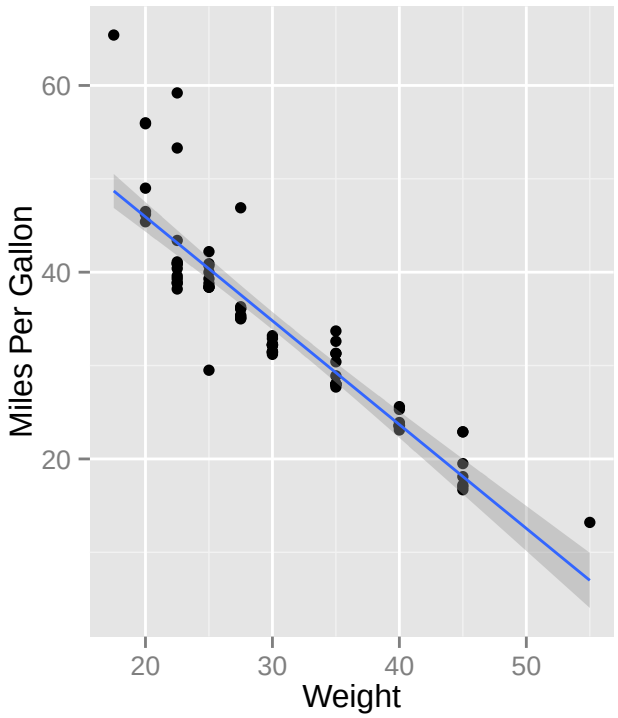
2 Gas Mileage and Automobile Weight

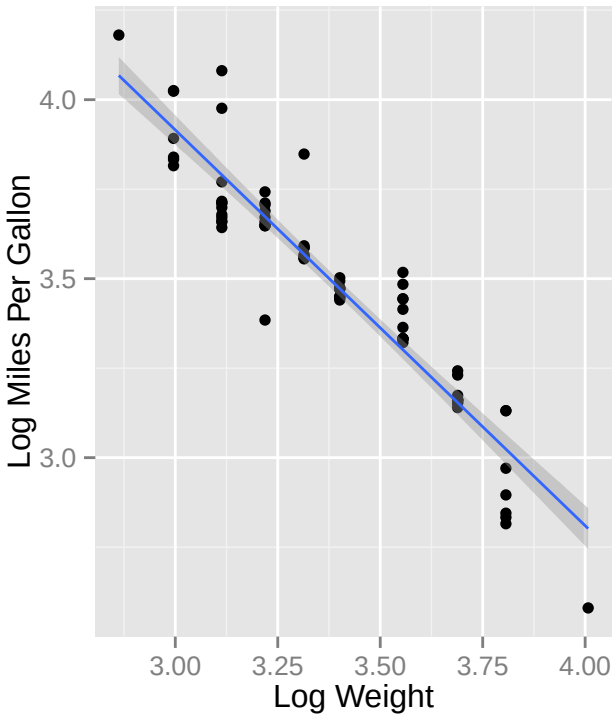
2(a)

```
# plot mpg = f(wt)
p2a <- ggplot(cars, aes(x = wt, y = mpg))
p2a <- p2a + geom_point()
p2a <- p2a + labs(y = "Miles Per Gallon", x = "Weight")
p2a <- p2a + geom_smooth(method = lm)

p2b <- ggplot(cars, aes(x = log_wt, y = log_mpg))
p2b <- p2b + geom_point()
p2b <- p2b + labs(y = "Log Miles Per Gallon", x = "Log Weight")
p2b <- p2b + geom_smooth(method = lm)

grid.arrange(p2a, p2b, ncol = 2)
```



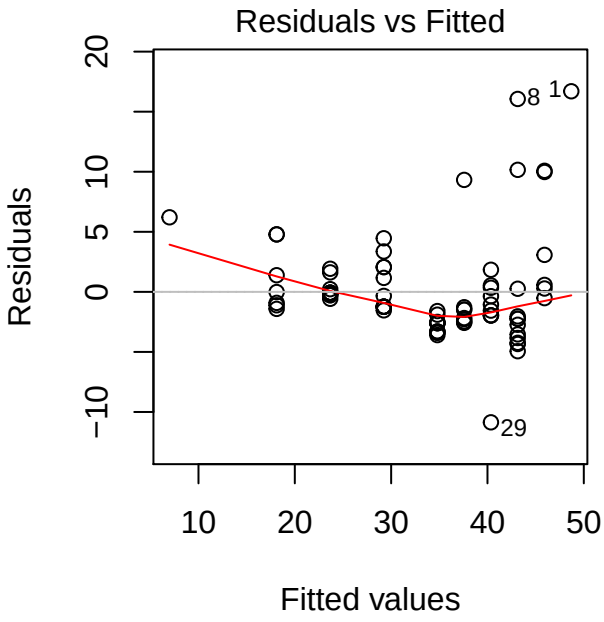


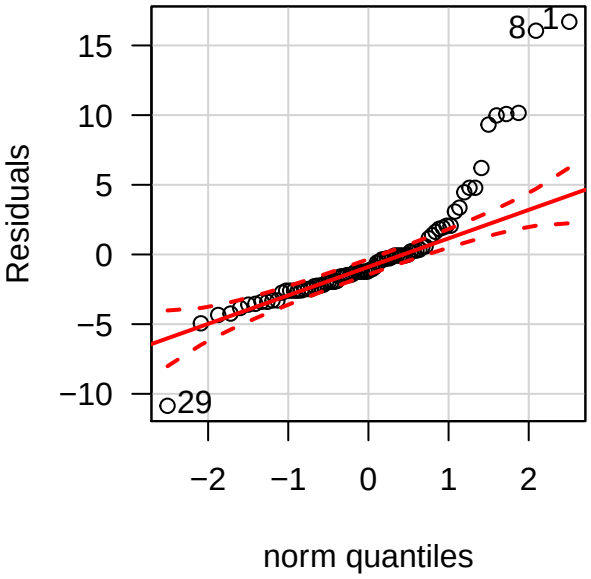
Based on these plots, little to no improvement in a linear fit would be gained from a log transformation of the data. The linear model appears to fit both data sets equally well.

2(b)

```
# Weight vs Mileage
lm.mpg.wt <- lm(mpg ~ wt, data = cars)
sum.lm.mpg.wt <- summary(lm.mpg.wt)
summary(lm.mpg.wt)
##
## Call:
## lm(formula = mpg ~ wt, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8601  -2.2698  -1.1768   0.4899  16.6983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.16545    1.86695   36.51   <2e-16 ***
## wt          -1.11222    0.05842  -19.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.281 on 80 degrees of freedom
## Multiple R-squared:  0.8192,Adjusted R-squared:  0.8169
## F-statistic: 362.4 on 1 and 80 DF,  p-value: < 2.2e-16
par(mfrow = c(1,2))
plot(lm.mpg.wt, which = 1)
abline(h = 0, col = "gray75")

qqPlot(lm.mpg.wt$residuals, las = 1, id.n = 3,
       ylab = "Residuals")
```






```
##  1  8 29
## 82 81  1
```

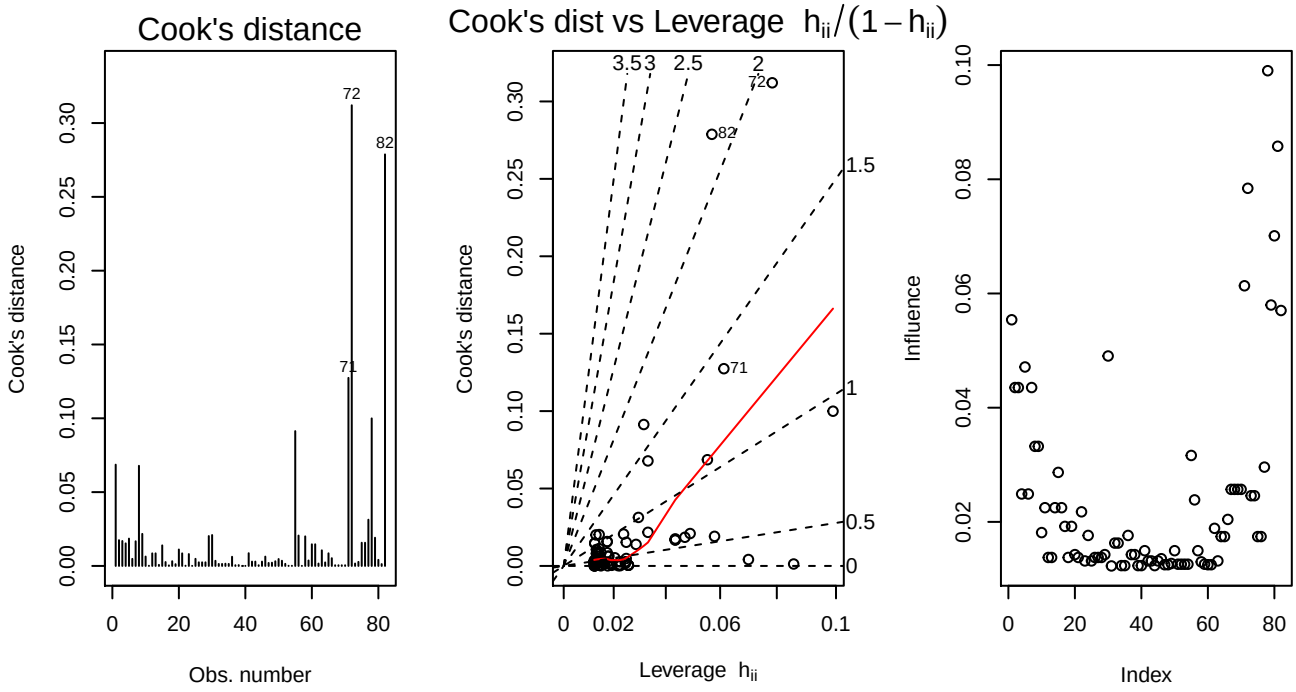
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	68.165	1.867	36.512	0.000
wt	-1.112	0.058	-19.037	0.000

```
shapiro.test(sum.lm.mpg.wt$residuals)
##
##  Shapiro-Wilk normality test
##
## data:  sum.lm.mpg.wt$residuals
## W = 0.8019, p-value = 3.94e-09
library(nortest)
ad.test(sum.lm.mpg.wt$residuals)
##
##  Anderson-Darling normality test
##
## data:  sum.lm.mpg.wt$residuals
## A = 5.443, p-value = 1.523e-13
```

Based on both the Shapiro-Wilks and Anderson-Darling tests for normality and the Q-Q plot, the residuals are not normally distributed. Also, the residual variance appears to increase in relationship to vehicle weight.

2(c)

```
par(mfrow=c(1,3))
plot(lm.logmpg.loghp,which = c(4,6))
plot(influence(lm.logmpg.loghp)$hat, ylab = "Influence")
```



Based on the leverage cuttoff and Cook's distance, no single point appears to have undue influence on the regression model.

2(d)

A weighted least squares model could be implemented to address the nonconstant variance of the residuals. Also, a fit to a second order polynomial could be used instead of a straight-line model. This seems reasonable, because the straight-line fit predicts that when you have a vehicle of zero weight the mileage will be 68 mpg, where intuitively I would interpret the mileage would assymptote to infinitely as vehicle weight approaches zero.