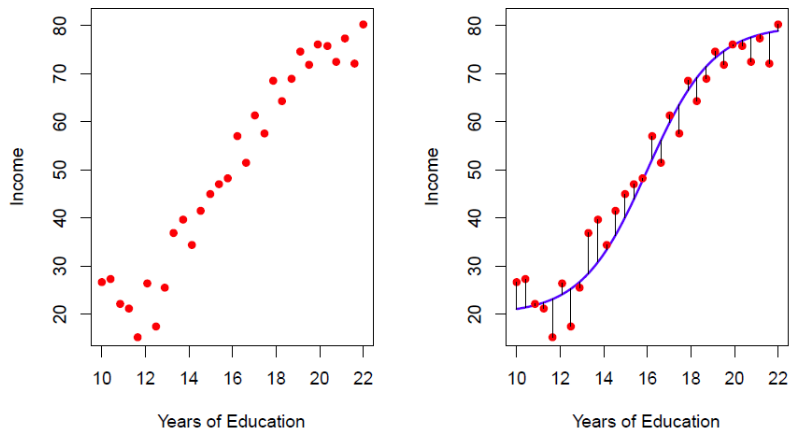
STA 325: Homework 1 Solutions

DUE: 11:59pm, September 13 (on Sakai)
COVERAGE: ISL Chapters 2–4

1. **[15 points]** Consider the following non-linear regression fit on annual income (in \$10,000) vs. years of education for $n = 30$ individuals:

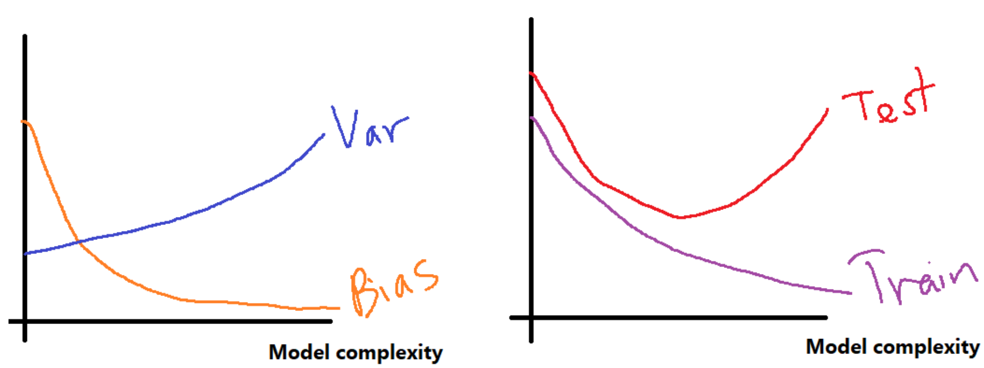


(a) Does the model fit the data well? Justify why or why not.

[3 points] Either YES or NO is fine, as long as they justify their answer. For YES, one reason may be that the errors are centered around zero with little autocorrelation. For NO, one reason may be that there's a sequence of points from $x = 14$ to $x = 17$ which are positive (but of course this could be random error).

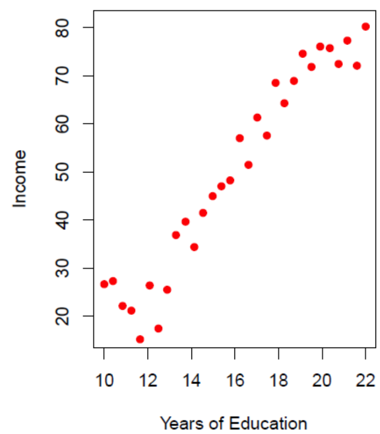
(b) Plot out what the bias, variance, test MSE and training MSE curves may look like as a function of model flexibility. Justify important features in these curves.

[6 points]



Bias curve should start high and decrease rapidly to zero. Variance curve should begin small (but non-zero) and increase gradually, since low irreducible error. Test curve should be U-shaped, with minimum at moderate complexity. Training curve should be decreasing and lower than test curve.

(c) Draw out below what the fitted model $\hat{f}(\cdot)$ may look like if we assumed high model flexibility. Use this to justify the test and training MSEs in part (b).



[3 points] The fitted function should "connect the dots" on the data points, resulting in zero training error but high test error.

(d) In the first plot, the fitted model $\hat{f}(\cdot)$ suggests significant slope changes at $x = 12$ and $x = 18$. Interpret what this means in terms of the problem. Based on purely income considerations, what advice would you give a graduating high-school student?

[3 points] Income increases at a more rapid rate between $x = 12$ and $x = 18$ years of education, and increases at a slower rate when $x < 12$ or $x > 18$. Based on purely income considerations, you should suggest the student to only get a bachelor's and master's degree (haven't we all heard this before?).

2. **[10 points]** For the classification problem (with K classes), we typically adopt the following conditional class probabilities:

$$p_k(x) = P[Y(x) = k], \quad k = 1, \cdots, K,$$

where $Y(x)$ is the discrete response at input predictors x. We discussed in-class the misclassification error measure:

$$\text{MCE}(x) := P[Y(x) \neq \hat{C}(x)]$$

where $\hat{C}(\cdot)$ is a chosen classifier function.

(a) Of the variables $Y(x)$, x and $\hat{C}(x)$, which are random? Which are not?

[2 points] $Y(x)$ is random, x and $\hat{C}(x)$ are not.

(b) In class, it was claimed that if $p_k(x)$ is known for each k and x, then the optimal (or "Bayes-optimal") classifier which minimizes $\text{MCE}(x)$ is:

$$\hat{C}(x) = k^*, \quad k^* := \underset{k=1,\cdots,K}{\text{argmax}}\ p_k(x).$$

Argue why this is true in words (or give a simple proof), justifying each step.

[4 points] Proof by contradiction. For a given x, suppose $\hat{C}(x) = k' \neq k^*$ minimizes $\text{MCE}(x)$. Then:

$$P[Y(x) \neq k'] = 1 - p_{k'}(x) > 1 - p_{k^*}(x) = P[Y(x) \neq k^*]$$

which contradicts the fact that $\hat{C}(x) = k'$ minimizes $\text{MCE}(x)$. Hence, $\hat{C}(x) = k^*$ is the Bayes-optimal classifier.

Note: it's fine to make the above argument in words.

(c) Explain the intuition behind this classifier in layman's terms (i.e., to someone who is not well-versed in statistics).

[2 points] If we know how the true probabilities of the response at each input, then the optimal classifier is the category with greatest probability of occurring.

(d) Why is this predictor not that useful in practice?

[2 points] We don't know the distribution of the response (i.e., the conditional class probabilities $p_k(x)$) in practice! We need to estimate this with data first.

3. **[10 points]** Suppose you are interested in predicting the number of hours spent on homework by freshmen and seniors. Let the predictor $x = 0$ for freshmen, and $x = 1$ for seniors. Your regression model is $Y(x) = \beta_0 + \beta_1 x + \epsilon$.

   (a) What is the interpretation of β_0?

   [2 points] The average number of hours spent by freshmen on homework.

   (b) What is the interpretation of $\beta_0 + \beta_1$?

   [2 points] The average number of hours spent by seniors on homework.

   (c) What is the interpretation of β_1?

   [3 points] The difference in average hours spent by seniors over freshmen on homework.

   (d) Do you expect the R^2 value for this model to be small or large? Why or why not?

   [3 points] I would expect the R^2 to be quite small, since it may not be a very informative predictor. To increase R^2, one may want to include more informative predictors, e.g., classes taken, major, etc.

4. **[10 points]** In logistic regression, we model the probability $p(x) = P[Y(x) = 1]$ as:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x.$$

(a) Solve the above equation to get an expression for $p(x)$.

[5 points] All work needs to be shown for full marks.

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

(b) Is there a linear relationship between x and $p(x)$?

[2 points] No, since a one-unit increase in x does not give a constant change in $p(x)$.

(c) How do we interpret the effect of a one-unit increase in x on the probability $p(x)$?

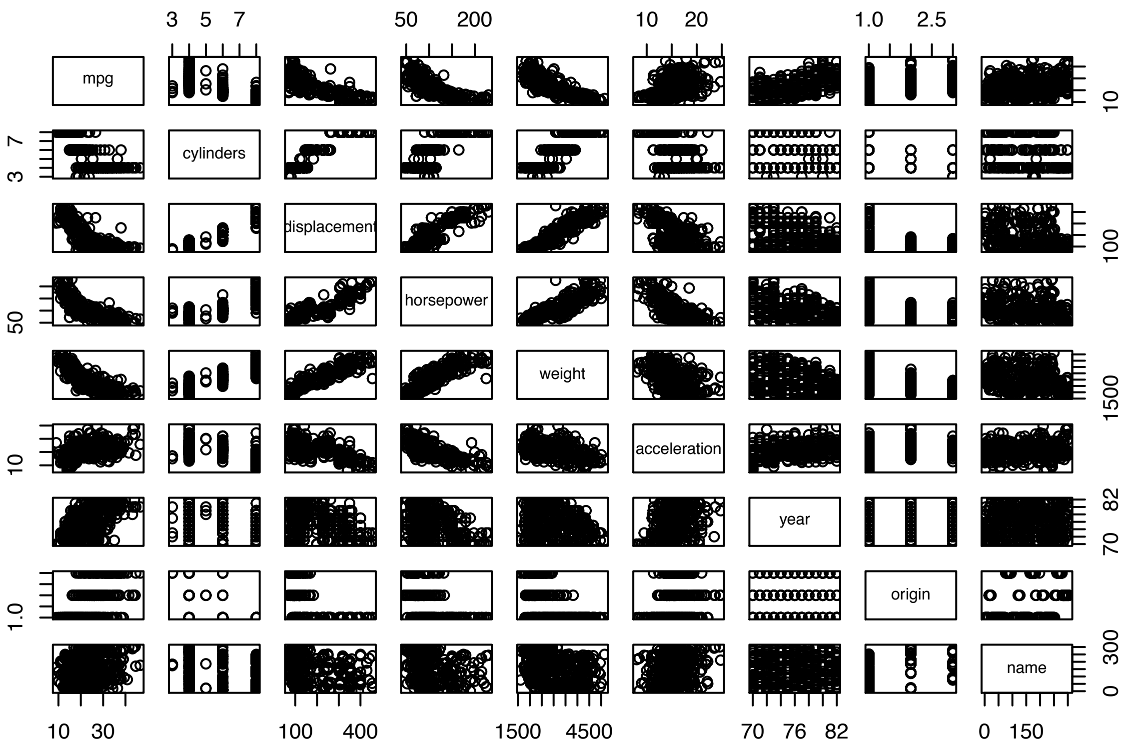
[3 points] A one-unit increase in x results in an increase of β for the log-odds of $p(x)$. Similarly, a one-unit increase in x results in the odds of $p(x)$ increasing by a factor of $\exp(\beta_1)$.

ISL Chapter 3, Exercise 9

(a)

```r
library(ISLR)

pairs(Auto)
```



(b)

```r
cor(subset(Auto, select=-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)

```r
lm1 = lm(mpg ~ . -name, data=Auto)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## 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
## 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)

Yes, there is a relatioship between the predictors and the response. We can show this by testing the null hypothesis that all the regression coefficients are zero. The F-statistic is far away from 1 (corresponding to a small p-value), showing evidence against the null hypothesis.

(ii)

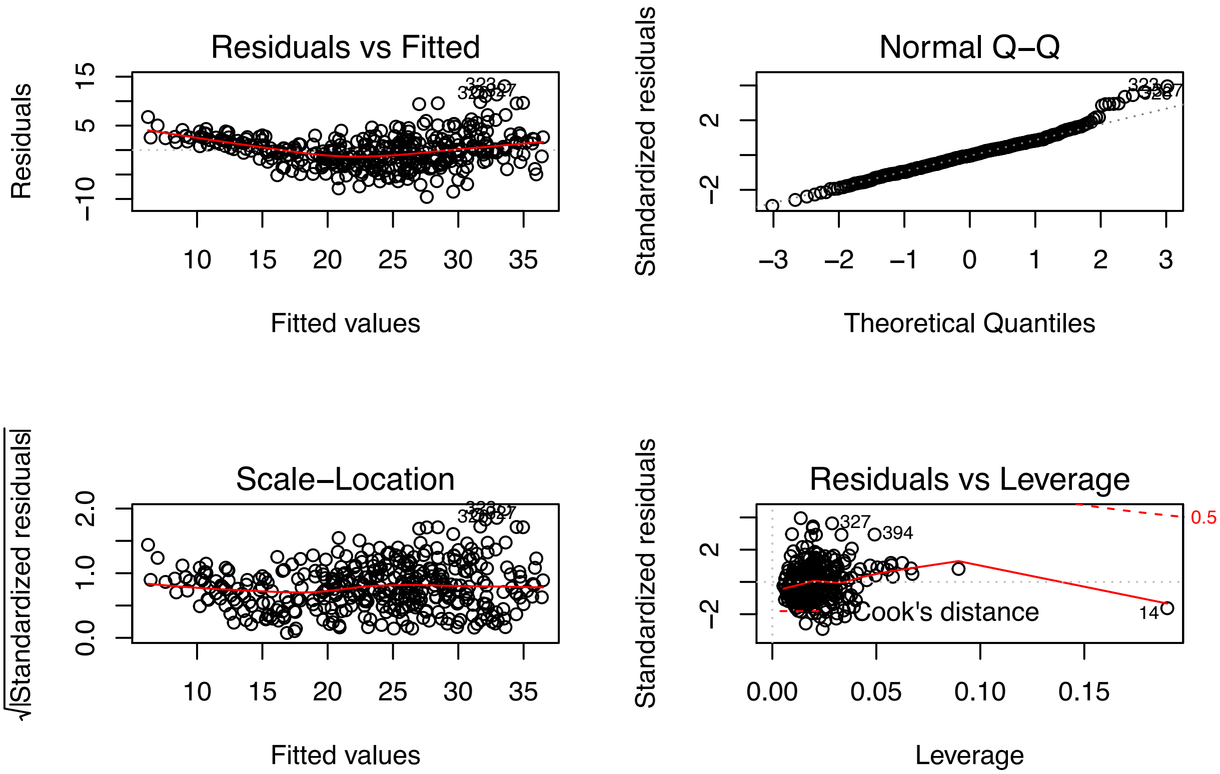
Based on the p-values of each predictor, we see that displacement, weight, year, and origin have a statistically significant relationship with the response, while cylinders, horsepower, and acceleration do not.

(iii)

The regression coefficient for year, 0.7508, suggests that for every additional year, MPG increases by 0.7508. In other words, cars become more fuel efficient by almost 1 mpg / year.

(d)

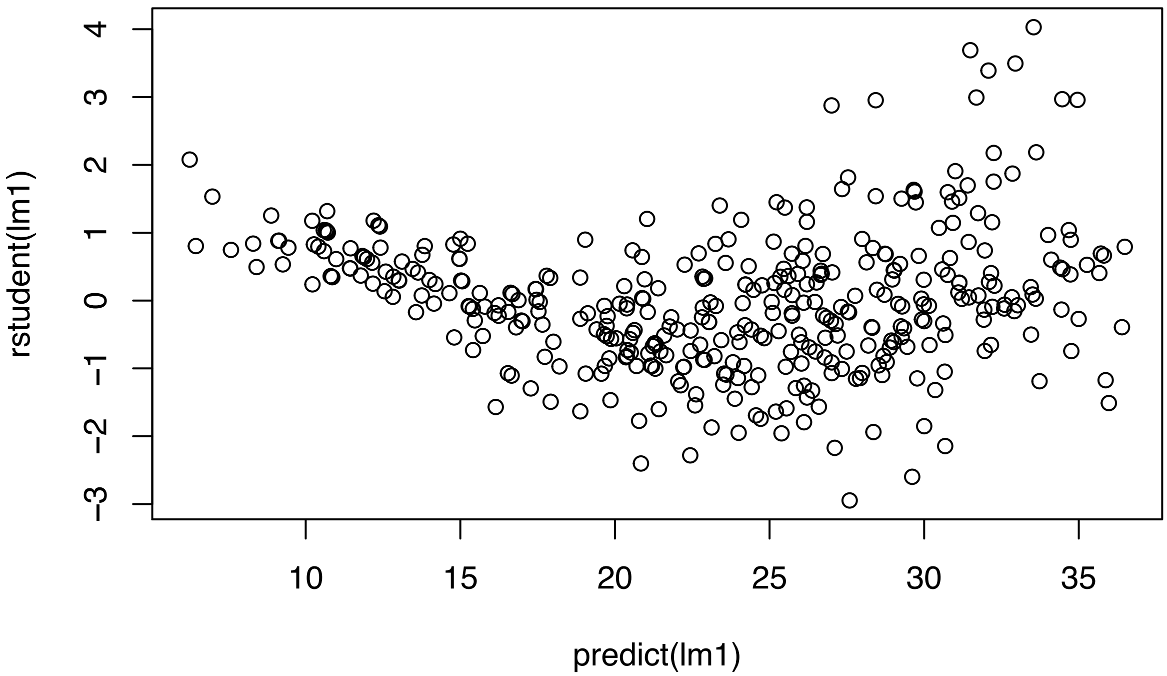
```r
par(mfrow=c(2,2))
plot(lm1)
```



The fit does not appear to be very good, since there is a noticeable curved pattern in the residual plots. From the leverage plot, point 14 appears to have high leverage, although it does not not a high magnitude residual.

There are also possible outliers, as can be observed in the plot of studentized residuals below (since some points have studentized residuals >3, which is a rough guideline).

```r
plot(predict(lm1), rstudent(lm1))
```



(e)

```
lm2 = lm(mpg ~ cylinders*displacement + displacement*weight, data=Auto)
summary(lm2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders               7.606e-01  7.669e-01   0.992    0.322
## displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

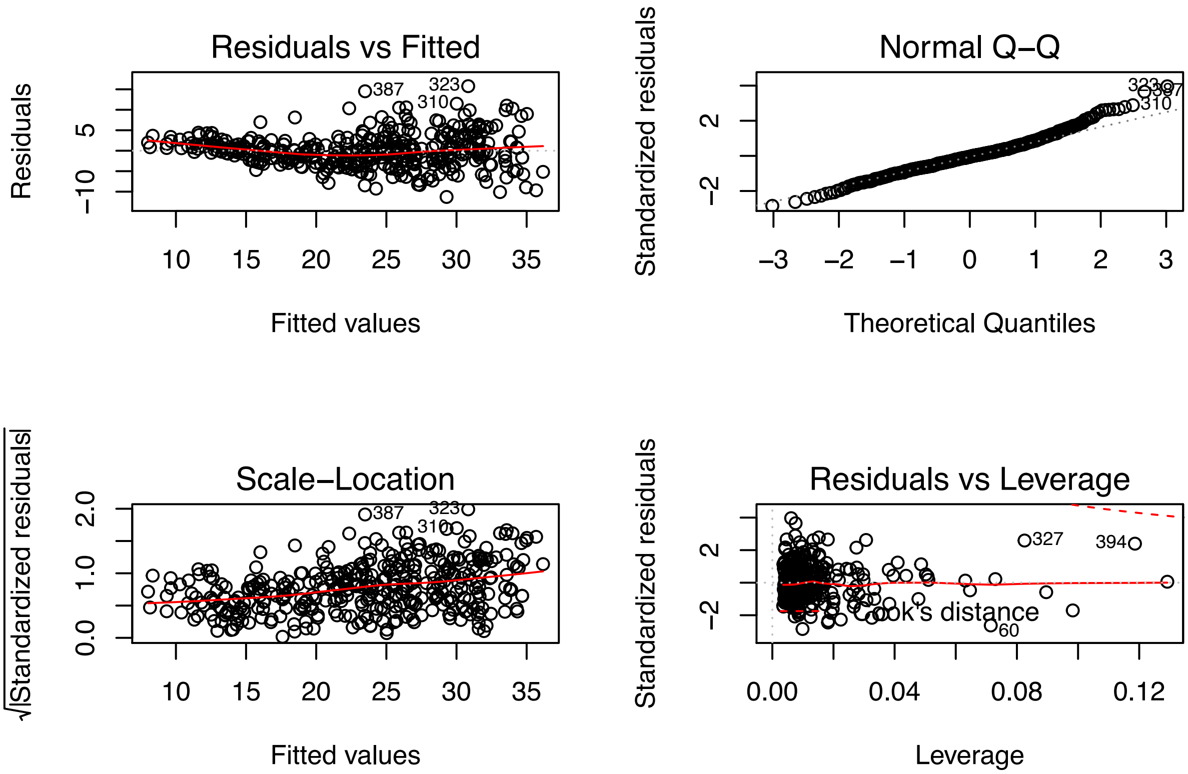
Two possible interaction effects have been added to the model. From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interaction between cylinders and displacement is not. Note: make sure that your interactions follow the hierarchy principle!

(f)

```
lm3 = lm(mpg ~ log(weight) + sqrt(horsepower) + acceleration + I(acceleration^2), data=Auto)
summary(lm3)
```

```
##
## Call:
## lm(formula = mpg ~ log(weight) + sqrt(horsepower) + acceleration +
##     I(acceleration^2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2932  -2.5082  -0.2237   2.0237  15.7650
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       178.30303   10.80451  16.503  < 2e-16 ***
## log(weight)       -14.74259    1.73994  -8.473 5.06e-16 ***
## sqrt(horsepower)   -1.85192    0.36005  -5.144 4.29e-07 ***
## acceleration       -2.19890    0.63903  -3.441 0.000643 ***
## I(acceleration^2)   0.06139    0.01857   3.305 0.001037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.99 on 387 degrees of freedom
## Multiple R-squared:  0.7414, Adjusted R-squared:  0.7387
## F-statistic: 277.3 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm3)
```



From the p-values, the log(weight), sqrt(horsepower), and acceleration^2 all have statistical significance of some sort. The residual plot has less of a discernible pattern than the plot with only linear regression terms. The leverage plot indicates a number of points with high leverage.

However, two problems are observed from the above plots: 1) the residuals vs fitted plot indicates heteroscedasticity (non-constant variance) in the model. 2) The QQ plot indicates non-normality of the residuals.

So, a better transformation should to be applied to our model. From the plot in 9(a), displacement, horsepower and weight show a similar nonlinear pattern against the response. This nonlinear pattern is close to a log form. So, we will use log(mpg) as our response variable.

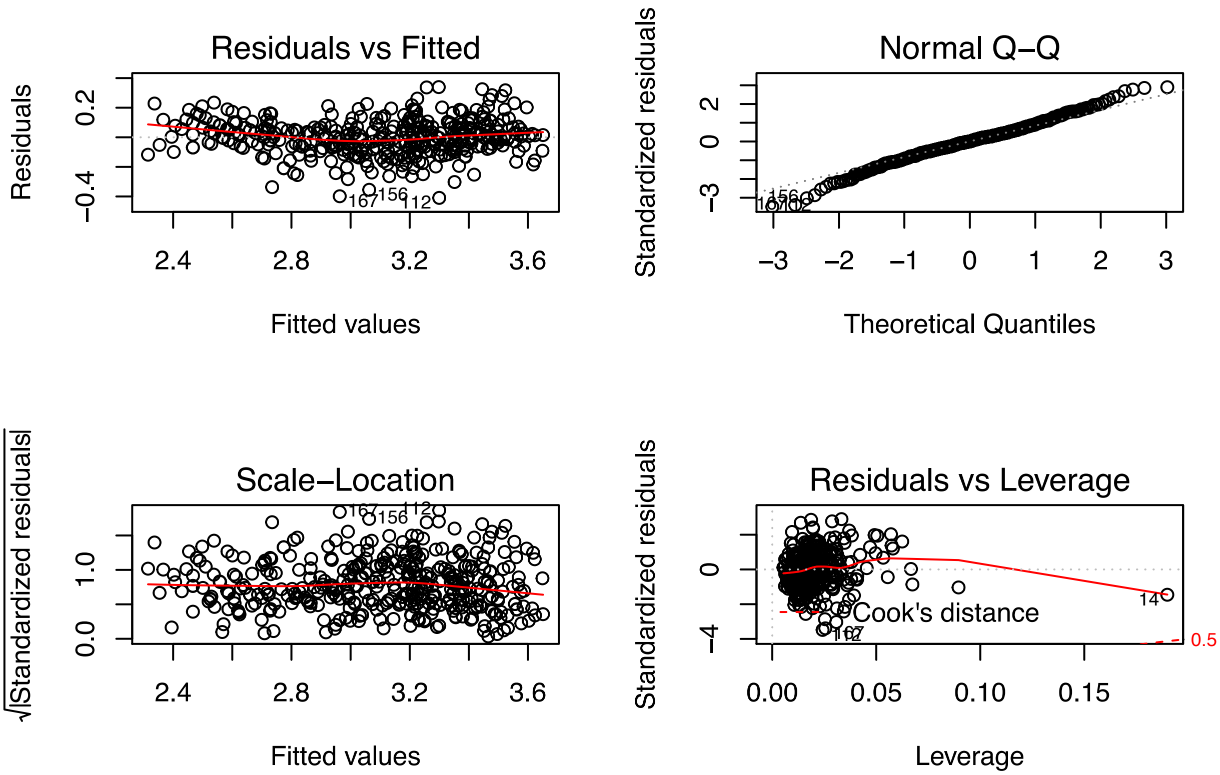
The output shows that log transform of mpg yields a better model fit (better R^2 and normality of residuals).

```
lm3<-lm(log(mpg) ~ cylinders + displacement + horsepower + weight +
          acceleration + year + origin, data=Auto)
summary(lm3)
```

```
##
## Call:
## lm(formula = log(mpg) ~ cylinders + displacement + horsepower +
##     weight + acceleration + year + origin, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40955 -0.06533  0.00079  0.06785  0.33925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.751e+00  1.662e-01  10.533  < 2e-16 ***
## cylinders    -2.795e-02  1.157e-02  -2.415  0.01619 *
## displacement  6.362e-04  2.690e-04   2.365  0.01852 *
```

```
## horsepower   -1.475e-03  4.935e-04  -2.989  0.00298 **
## weight       -2.551e-04  2.334e-05 -10.931  < 2e-16 ***
## acceleration -1.348e-03  3.538e-03  -0.381  0.70339
## year          2.958e-02  1.824e-03  16.211  < 2e-16 ***
## origin        4.071e-02  9.955e-03   4.089 5.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1191 on 384 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8773
## F-statistic: 400.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(lm3)
```



Note: there are many other correct ways to have done this problem.

ISL Chapter 3, Exercise 14

(a)

```r
set.seed(1)
x1 = runif(100)
x2 = 0.5 * x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

$$Y = 2 + 2X_1 + 0.3X_2 + \epsilon$$

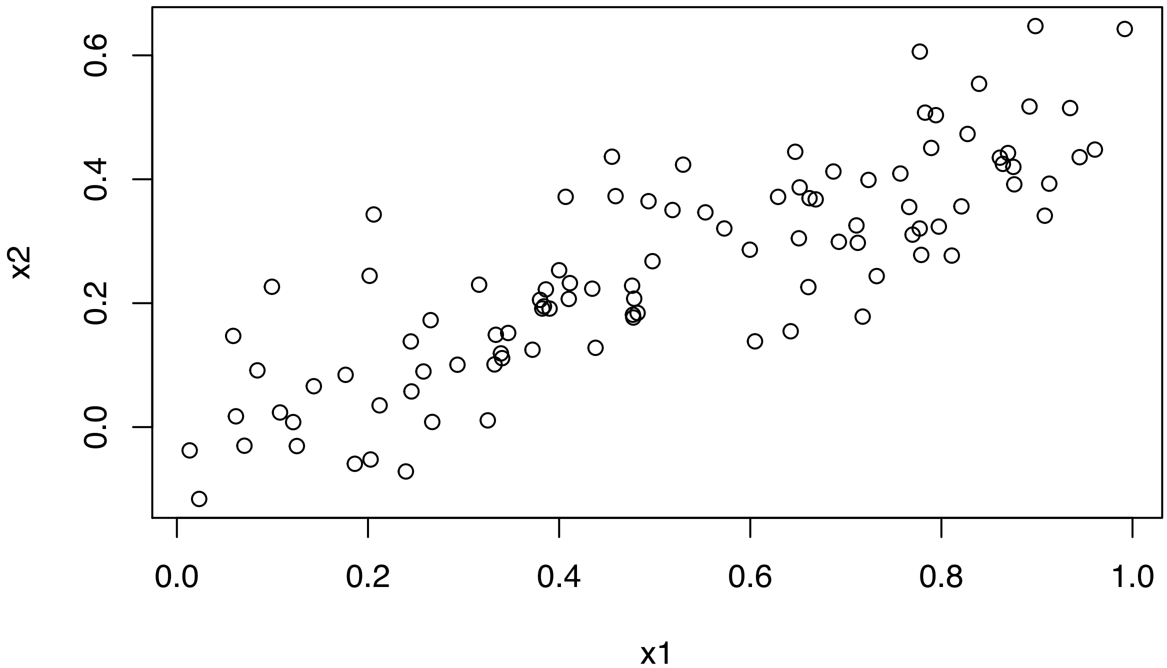
$$\beta_0 = 2, \beta_1 = 2, \beta_3 = 0.3$$

(b)

```r
cor(x1, x2)
```

```
## [1] 0.8351212
```

```r
plot(x1, x2)
```



(c)

```r
lm.fit = lm(y ~ x1 + x2)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

The regression coefficients are close to the true coefficients, although with high standard error. We can reject the null hypothesis for β_1 because its p-value is below 0.05. We cannot reject the null hypothesis for β_2 because its p-value is well above the typical 0.05 cutoff.

(d)

```r
lm.fit = lm(y ~ x1)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

Yes, we can reject the null hypothesis for the regression coefficient since the p-value for its t-statistic is near zero.

(e)

```r
lm.fit = lm(y ~ x2)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

Yes, we can reject the null hypothesis for the regression coefficient since the p-value for its t-statistic is near zero.

(f)

No. Because x1 and x2 have collinearity, it is hard to distinguish their effects when they are both in the model. When they are included separately, the linear relationship between y and each predictor is seen more clearly. In other words, parts (c) and (e) are testing different things: part (c) test whether x_2 is influential marginal on x_1, while part (e) tests whether x_2 is influential at fixed levels of x_1.

7. [**5 points**] ISL Chapter 4, Exercise 8.

Note that 1-NN has a training error rate of 0%, since the nearest neighbour of a training observation is itself. Since the data is split 50-50 into training and testing, it follows that the test error rate for the fitted 1-NN classifier is 2*18% = 36%. The fitted logistic regression classifier has a lower test error rate of 30%, so it should be preferred over 1-NN.

8. **[BONUS: 10 points]** Assume the general learning model:

$$Y(x) = f(x) + \epsilon,$$

where $Y(x)$ is the response at input predictors x, and ϵ is a random error term. Instead of the MSPE discussed in class, suppose we use a different error measure – the mean absolute predictive error (MAPE):

$$\text{MAPE}(x) := \mathbb{E}[|Y(x) - \hat{f}(x)|].$$

We wish to find the optimal predictor under this new MAPE error measure.

(a) Of the variables $Y(x)$, x, ϵ and $\hat{f}(x)$, which are random? Which are not?
[1 point] $Y(x)$ and ϵ are random, x and $\hat{f}(x)$ are not.

(b) Let Z be a continuous random variable with distribution function $F(\cdot)$. For any number m, show that:

$$\mathbb{E}[|Z - m|] = \int_{-\infty}^{m} (m - z)\, dF(z) + \int_{m}^{\infty} (z - m)\, dF(z).$$

[1 point] Self-explanatory.

(c) Define $m^* = \text{med}(Z)$ as the *median* of Z, satisfying $F(m^*) = 0.5$. Using (b), show that for any number m greater than m^*, we have:

$$\mathbb{E}[|Z - m|] - \mathbb{E}[|Z - m^*|] = (m - m^*)\,[P(Z \leq m^*) - P(Z > m^*)] + 2\int_{m^*}^{m} (m - z)\, dF(z).$$

[3 points] Using part (b), we get:

$$\begin{aligned}
\mathbb{E}[|Z - m|] &= \int_{-\infty}^{m^*} (m - z)dF(z) + \int_{m^*}^{m} (m - z)dF(z) + \int_{m}^{\infty} (z - m)dF(z) \\
\mathbb{E}[|Z - m^*|] &= \int_{-\infty}^{m^*} (m^* - z)dF(z) + \int_{m^*}^{m} (z - m^*)dF(z) + \int_{m}^{\infty} (z - m^*)dF(z) \\
&= \int_{-\infty}^{m^*} (m^* - z)dF(z) + \int_{m^*}^{m} (m - m^*)dF(z) \\
&\qquad - \int_{m^*}^{m} (m - z)dF(z) + \int_{m}^{\infty} (z - m^*)dF(z).
\end{aligned}$$

Hence:

$$\begin{aligned}
\mathbb{E}[|Z-m|]-\mathbb{E}[|Z-m^*|] &= (m-m^*)P(Z\le m^*)-(m-m^*)P(m^*<Z\le m) \\
&\quad +(m^*-m)P(Z>m)+2\int_{m^*}^{m}(m-z)dF(z) \\
&= (m-m^*)P(Z\le m^*)-(m-m^*)P(Z>m^*) \\
&\quad +2\int_{m^*}^{m}(m-z)dF(z),
\end{aligned}$$

which proves the claim.

(d) Using (c), argue that $\mathbb{E}[|Z-m|]-\mathbb{E}[|Z-m^*|]\ge 0$ for any $m>m^*$.

[1 point] Self-explanatory.

(e) Using (d), show that if $Y(x)$ is known for each x, the optimal predictor minimizing MAPE(x) is $\hat{f}(x)=\text{med}[Y(x)]$.

[1 point] Argue the same inequality as in (c) holds for $m<m^*$. Since $\mathbb{E}[|Z-m|]-\mathbb{E}[|Z-m^*|]\ge 0$ for any m, and $m=m^*$ achieves this inequality exactly, it follows that $m=m^*$ minimizes $\mathbb{E}[|Z-m|]$. The result is proved by setting $Z=Y(x)$.

(f) Explain the intuition behind this predictor in layman's terms (i.e., to someone who is not well-versed in statistics).

[1 point] If we know the distribution of the response, then the optimal predictor under MAPE is the median response at the input we wish to predict.

(g) From (e), the optimal predictor minimizing MSPE (i.e., $\hat{f}(x)=\mathbb{E}[Y(x)]$) is different from the optimal predictor minimizing MAPE (i.e., $\hat{f}(x)=\text{med}[Y(x)]$). Give a real-world scenario where the latter predictor may be more preferable to the former.

[2 points] This is an interesting question. There are a few possible scenarios:

- The noise term ϵ does not have finite variance, i.e., $\text{Var}(\epsilon)=\infty$. This happens when the noise random variable ϵ has very heavy tails. Practically, if we adopt the standard approach of modeling $\mathbb{E}[Y(x)]$, then we may end up with spurious and useless prediction results, as well as huge predictive intervals.
- The noise term ϵ is heavily left or right skewed. In such a scenario, the standard approach of modeling $\mathbb{E}[Y(x)]$ may yield a overly complex model which has poor predictive power (high variance). If we instead model $\text{med}[Y(x)]$, we get a more robust and likely simpler model which improves predictive power.
- Presence of outliers in data. By modeling the median $\text{med}[Y(x)]$ instead of the mean $\mathbb{E}[Y(x)]$, the fitted model will be much more resistant to the effect of outliers data points. This is because the median is a more robust measure of centrality compared to the mean.