Homework 2

Instructions

Due: 5:00pm on Wednesday, September 22nd

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Theory

Problem 1

Based on ISLR Exercise 3.1

Write out the null hypotheses to which the p-values given in Table 3.4 (p. 75 ISLR) correspond. Explain what conclusions you can draw based on these particular p-values.

Let $\beta_0, \beta_1, \beta_2, \beta_3$ be the coefficient parameters on Intercept, TV, radio, and newspaper, respectively. Then the 4 p-values correspond to the hypotheses:

1. $H_0 : \beta_0 = 0$
2. $H_0 : \beta_1 = 0$
3. $H_0 : \beta_2 = 0$
4. $H_0 : \beta_3 = 0$

In particular, each tests the claim that the respective coefficient is 0 in the model based on all 3 variables (this is distinct from testing with the associated coefficient is 0 in the corresponding SLR model).

Since the p-values for the first three tests were small (<0.0001), can reject the corresponding null hypotheses in favor of the alternative: the data suggests a linear relationships between sales and TV, as well as between sales and radio,in the presence of other variables. The data also suggests that expected sales will not be 0 when all predictors are equal to 0. On the other hand, the data suggests there may be no linear relationship between sales and newspaper, *in the presence of TV and radio.*

Problem 2

Based on ISLR Exercises 3.5 and 3.6

(a) Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y}_i = x_i\hat{\beta}$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{j=1}^{n} x_j^2\right)$$

Show that we can write

$$\hat{y}_i = \sum_{j=1}^{n} a_j y_j$$

for some constants a_j (that might depend on the $x_1, \ldots, x_n$, but that do not depend on any of $y_1, \ldots, y_n$.) Give the explicit formula for a_j.

(b) Use equation 3.4 in the text to show that for SLR, the least squares line always passes through the point $(\bar{x}, \bar{y})$

(c) **Optional, does not need to be submitted** Show that the R^2 statistic in formula (3.17) is equal to the square of the correlation betwene X and Y, as given in formula (3.18). For simplicitly, you may assume that $\bar{x} = \bar{y}$ (although the result is true more generally as well).

(a) Here, we evaluate the formula for $\hat{y}$ using the formula for $\hat{b}$. Let $s^2 = \left(\sum_{j=1}^{n} x_j^2\right)$.

$$\begin{aligned}
\hat{y}_i =& x_i \hat{\beta} \\
=& x_i \frac{\left(\sum_{j=1}^{n} x_j y_j\right)}{s^2} \\
=& \frac{\left(\sum_{j=1}^{n} x_i x_j y_j\right)}{s^2} \\
=& \sum_{j=1}^{n} \frac{x_i x_j}{s^2} y_j \\
=& \sum_{j=1}^{n} a_j y_j
\end{aligned}$$

where

$$a_j = \frac{x_i x_j}{s^2}$$

(b) Observe that when the regression equation is evaluated at $x = \bar{x}$,

$$\begin{aligned}
\hat{y} =& \hat{\beta}_1 \bar{x} + \hat{\beta}_0 \\
=& \hat{\beta}_1 \bar{x} + \bar{y} - \hat{\beta}_1 \bar{x} \\
=& \bar{y}
\end{aligned}$$

which means that the regression line passes through the point $(\bar{x}, \bar{y})$.

Applied

Problem 3

1000 large seismic events around Fiji have been collected in a data set called `quakes` that is built into R. You can learn more about it with the following commands:

Earthquake detection Included in the data set is a column recording the number of stations that detected each earthquake. This refers to a global network of seismographs and it stands to reason that the larger the quake, the more widely it will be detected.

(a) Create a plot of the relationship between `stations` and `magnitude`. How would you characterize the relationship? (If you see overplotting, you may want to add `jitter` to your points or make them transparent by playing with the `alpha` value.)

(b) If there was actually *no relationship* between the two variables, what would you expect the slope of a linear model to be? What about the intercept?

(c) Fit a linear model for `stations` as a function of `mag` and record the regression coefficients. Add the corresponding least squares line to the plot from exercise 1. Interpret your slope and intercept in the context of the problem.

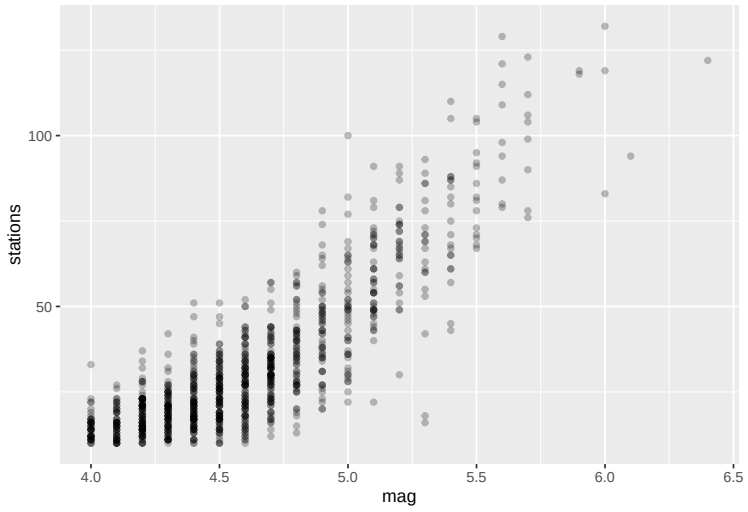
(d) Using formulas 3.8 and 3.9 on page 66 of ISLR , calculate a 95% confidence interval for the slope of the model that you fit in exercise 3 (you can use R as a calculator to assist with arithmetic). Confirm the calculation by applying the `confint()` function to your linear model.

(e) How many stations do you predict would be able to detect an earthquake of magnitude 7.0?

(f) Parts (a) - (e) in this problem involve elements of *data description*, *inference*, and/or *prediction*. Which was the dominant goal in each question?

(a) The scatterplot suggests a relatively strong positive linear relationship between magnitude and stations, with some noticably deviation from linearity for small magnitude earthquakes.

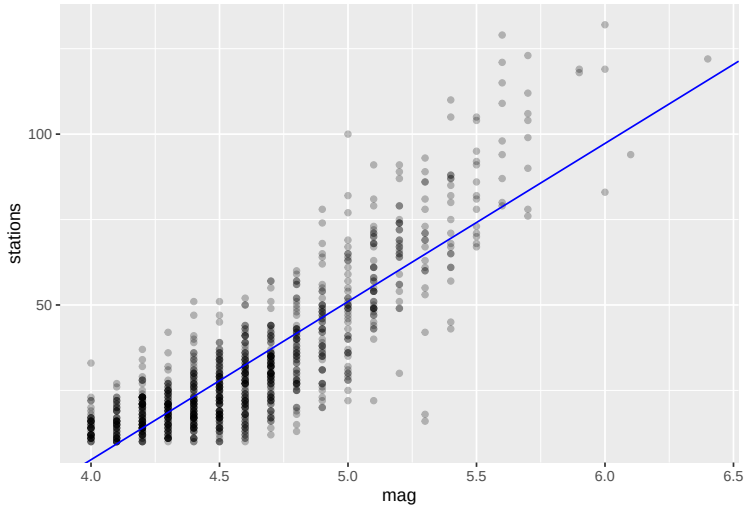
```
quakes %>% ggplot( aes( x = mag, y = stations)) + geom_point(alpha= 0.25)
```



(b) If there were truly no relationship between the two variables, we would expect the regression line to have slope of 0 and intercept equal to the average value of the response.

(c) The model suggests every unit increase in magnitude corresponds to an expected increase of 46.28 stations, and that -180 stations will detect a magnitude 0 earthquake (obviously absurd).

```
m1<-lm(stations ~ mag, data = quakes)
b0<- summary(m1)$coefficients[1]
b1<- summary(m1)$coefficients[2]
quakes %>% ggplot( aes( x = mag, y = stations)) +
  geom_point(alpha= 0.25) +
  geom_abline(intercept = b0, slope = b1, color = "blue")
```




```
#alternatively add layer geom_smooth(method = "lm", se = F)
```

(d) We use the data Residual Standard Error $RSE = 11.5$ as our estimate of σ, the standard deviation of errors, and note that

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```
RSE <- summary(m1)$sigma
sx <- sum( (quakes$mag - mean(quakes$mag) ) ^2)
SE_b1 <- sqrt(RSE^2 / sx)

data.frame(RSE,  SE_b1)
```

```
##        RSE     SE_b1
## 1 11.50061 0.9033955
```

Our 95% confidence interval has critical value approximately 1.96, and so formula

$$\hat{\beta}_1 \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}_1)^2$$

```
data.frame(Lower = b1 - 1.96*SE_b1, Upper = b1 + 1.96*SE_b1)
```

```
##      Lower    Upper
## 1 44.51156 48.05287
```

This is verified by the `confint` function:

```
confint(m1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -188.64628 -172.20238
## mag           44.50944   48.05498
```

(e) The linear model predicts 143.55 stations will detect a magnitude 7 earthquake.

```
predict(m1 , data.frame( mag = 7) )
```

```
##        1
## 143.5511
```

(f) *Data Description*: a

Inference: b, c, d

Prediction: c, e

Problem 4

One good way to assess whether your fitted model seems appropriate is to simulate data from it and see if it looks like the data that you observed. We'll do this for the `mag` and `station` data in the `quakes` data set.

(a) To begin, to generate data similar to `mag`, we will sample with replacement from the existing data (think of this like performing a bootstrap sample). Some observations will likely be selected more than once, These repeated values are standing as approximates for other similar, but unobserved values. Use the `sample()` function in R to create a bootstrap sample from `mag`, and save it as the vector `sim_mag`.

(b) We now need to theorize the functional relationship between `station` and `mag`. Since we fit a linear model previously in Problem (1), we can use that as a starting point. To generate the $\hat{y}$ predicted values from your linear function based on your simulated data in (a), we can define an R function. Replace the line beginning with # in the code chunk below with the formula you found in Problem 1 (i.e. something like `10 - 2x`)

```
f_hat <- function(x){
  # your formula for the linear function goes here
}
```

(c) Generate your predicted values by applying the `f_hat` function you just made to the vector of predictors `sim_mag` and store the result as the vector `pred_stations`.

(d) Now, we will simulate observed y's by adding random error to each predicted value. Estimate the standard deviation of the error using the observed RSE from your model in Problem 3 and store this value as the variable `obs_rse`.

(e) We will assume that errors are Normally distributed with mean of 0 and standard deviation of `obs_rse`. The function `rnorm(n, mean, sd)` generates n observations from a Normal distribution with mean `mean` and standard deviation `sd`. Use this function to generate 1000 independent errors.

(f) Create a vector of simulated observed values by adding together your vector of predicted values and the vector of errors, and save the new vector as `sim_stations`.

(g) Create a data frame of simulated data called `quakes_sim` by applying the `data.frame()` function to the vectors `sim_mag` and `sim_stations`.

(h) Perform exploratory data analysis on this simulated data set. How does your simulated data compare to the actual observed data? How might you change your simulation to make the data more consistent with the observed data?

(a)

```
set.seed(1010)
sim_mag <- sample(quakes$mag, replace = T)
```

(b) Based on work in Problem 3, the least squares regression equation is

$$\hat{f}(x) = -180.42 + 46.28x$$

```
f_hat <- function(x){
  b0 + b1*x
}
```

(c)

```
pred_stations <- f_hat(sim_mag)
```

(d) Based on the linear model, our rse is 11.49.

```
rse <- sd(m1$residuals)
rse
```

```
## [1] 11.49485
```

```
# alternatively, use
# rse <- summary(m1)$sigma
```

(e)

```
set.seed(919)
error <- rnorm(1000, 0, rse)
```

(f)

```
sim_stations <- pred_stations + error
```

(g)

```
quake_sim <- data.frame(sim_mag, sim_stations, pred_stations)
```

(h) The simulated data shows approximately the same strength of linear relationship as the actual data, as evidenced by correlation, R^2, and MSE. However, the scatterplots are far from identical:

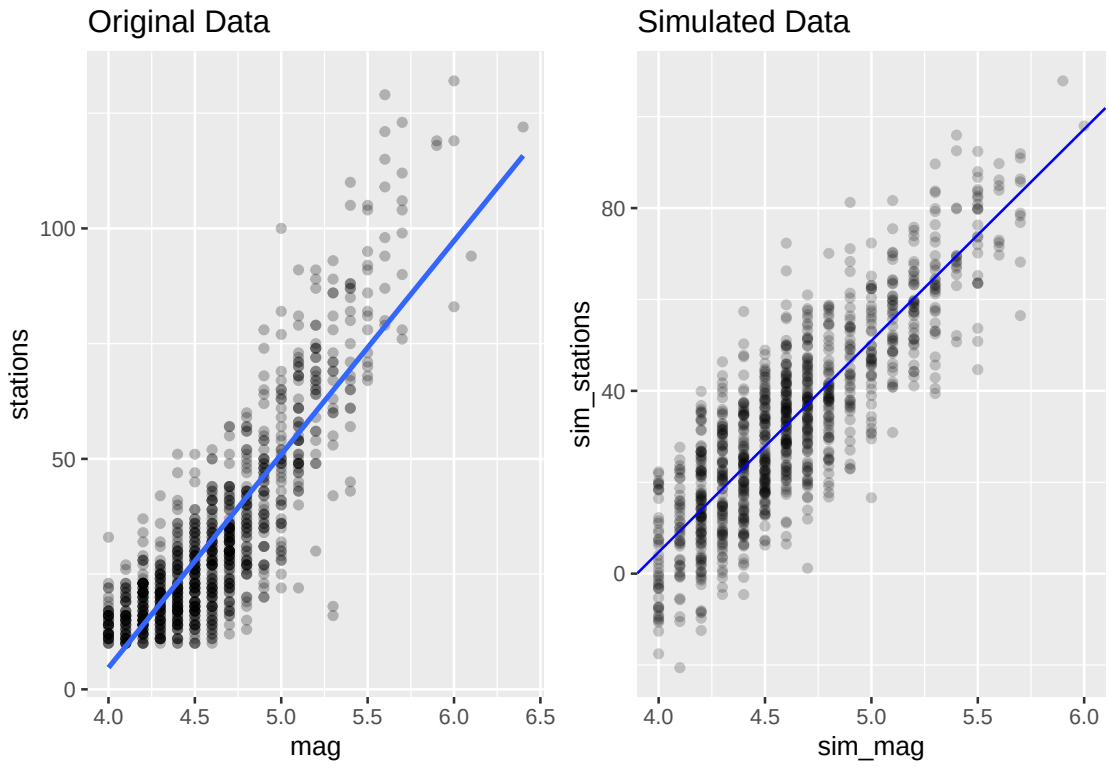
- The simulated data shows points concentrated relatively equally along the regression line. On the other hand, the actual data is concentrated in the lower left corner of the graph.
- The simulated data reports a negative number of stations for some quakes with low magnitude, while the real data has no such pattern.
- The simulated data has some signs of non-linearity in the lower left and upper right corners.
- The residuals in the actual data do not appear to have constant variance (residuals corresponding to lower magnitude have lower variance).

One fix for several of the above discrepancies is to use a more appropriate distribution for the simulated errors (for example, by resampling from the actual residuals, or by using domain knowledge to produce a theoretical distribution.)

```
library(gridExtra)
plot1 <- quakes %>% ggplot( aes( x = mag, y = stations)) +
  geom_point(alpha= 0.25) +
  geom_smooth(method = "lm", se = F)+
  labs(title = "Original Data")

plot2<-ggplot(quake_sim, aes(x = sim_mag, y = sim_stations))+
  geom_point(alpha=.2)+
  geom_abline(intercept = b0, slope = b1, color = "blue")+
  labs(title = "Simulated Data")

grid.arrange(plot1,plot2,nrow = 1)
```




```
sim_mod<-lm(sim_stations ~ sim_mag, data = quake_sim)
summary(sim_mod)
```

```
##
## Call:
## lm(formula = sim_stations ~ sim_mag, data = quake_sim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.546  -7.843  -0.200   8.229  40.093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -174.0868     4.1877  -41.57   <2e-16 ***
## sim_mag       44.8512     0.9013   49.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.38 on 998 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.7124
## F-statistic:  2476 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = stations ~ mag, data = quakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -48.871  -7.102  -0.474   6.783  50.244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -180.4243     4.1899  -43.06   <2e-16 ***
## mag           46.2822     0.9034   51.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.5 on 998 degrees of freedom
## Multiple R-squared:  0.7245, Adjusted R-squared:  0.7242
## F-statistic:  2625 on 1 and 998 DF,  p-value: < 2.2e-16
```

Problem 5

Based on ISLR Exercise 3.9

This question uses the `Auto` data set, loaded from the `ISLR` library, as well as the `ggpairs` function from the `GGally` library. Both libraries are loaded by running the code chunk below.

```
library(ISLR)
library(GGally)
```

You can learn more about tbe data set with the following commands:

(a) When the `ggpairs` function is applied to a data frame, it creates a matrix of pairwise scatterplots and correlations for all variables in the data frame (using `ggplot` styling conventions). Use this function to create pairwise scatterplots for all **quantitative** variables in the `Auto` data set. *You may want to adjust the displayed figure dimensions using chunk options (the gear in upper right of chunk)*

(b) Use the `lm` function to fit a MLR model with `mpg` as the response and all other quantitative variables as predictors. Then use the `summary` function to print the results.

(c) Based on your model, does there appear to be a relationship between the predictors and response? Which predictors have statistically significant relationship with the response? What does the coefficient for the `year` variable suggest? Justify your answers.

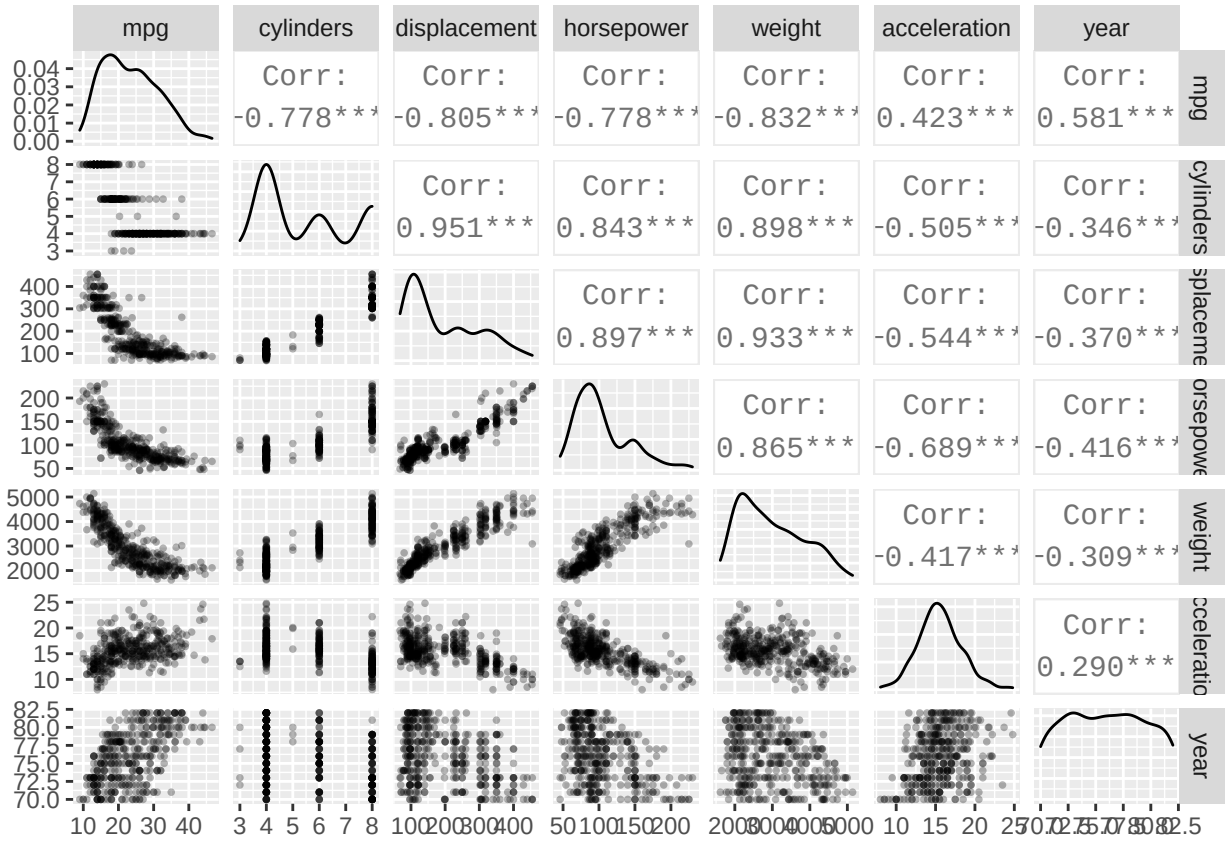
(d) Create diagnostic plots for the linear regression fit. Comment on any problems you observe. Do the residual plots suggest any unusually large outliers? Do leverage plots suggest any observations with unusually large leverage?

(e) Fit a linear regression model with at least 3 interaction terms of your choice. Do any of these interactions terms appear significant?

(f) Try two different transformations of two different variables. Comment on the effect.

(a) Unfortunately, with 49 plots arranged in a 7x7 grid, it is difficult to discern individual plot features for images output in .pdf file. However, the built-in R view can provide much larger resolutions allowing meaningful data exploration.

```
Auto_q<-Auto %>% select(-name, -origin)
ggpairs(Auto_q, aes(alpha = .15),
        lower = list(continuous = wrap("points", alpha = 0.3, size=1, shape =16))
)
```



Note that the ***origin*** *variable is actually categorical, even though it is stored in the data frame as a numeric variable. It should not be included in the scatterplots or linear models.*

(b)

```
mpg_mod<-lm(mpg ~., data = Auto_q)
summary(mpg_mod)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = Auto_q)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6927 -2.3864 -0.0801  2.0291 14.3607
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.454e+01  4.764e+00  -3.051  0.00244 **
## cylinders    -3.299e-01  3.321e-01  -0.993  0.32122
## displacement  7.678e-03  7.358e-03   1.044  0.29733
## horsepower   -3.914e-04  1.384e-02  -0.028  0.97745
## weight       -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
## acceleration  8.527e-02  1.020e-01   0.836  0.40383
## year          7.534e-01  5.262e-02  14.318  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.435 on 385 degrees of freedom
## Multiple R-squared:  0.8093, Adjusted R-squared:  0.8063
## F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

(c) The data does suggest a linear relationship between `mpg` and several of the predictors: the R^2 statistic suggests 81.82% of variation in `mpg` is explained by the linear model, and p-value of the F statistic suggests there is very strong evidence to reject the null hypothesis all coefficients are 0.

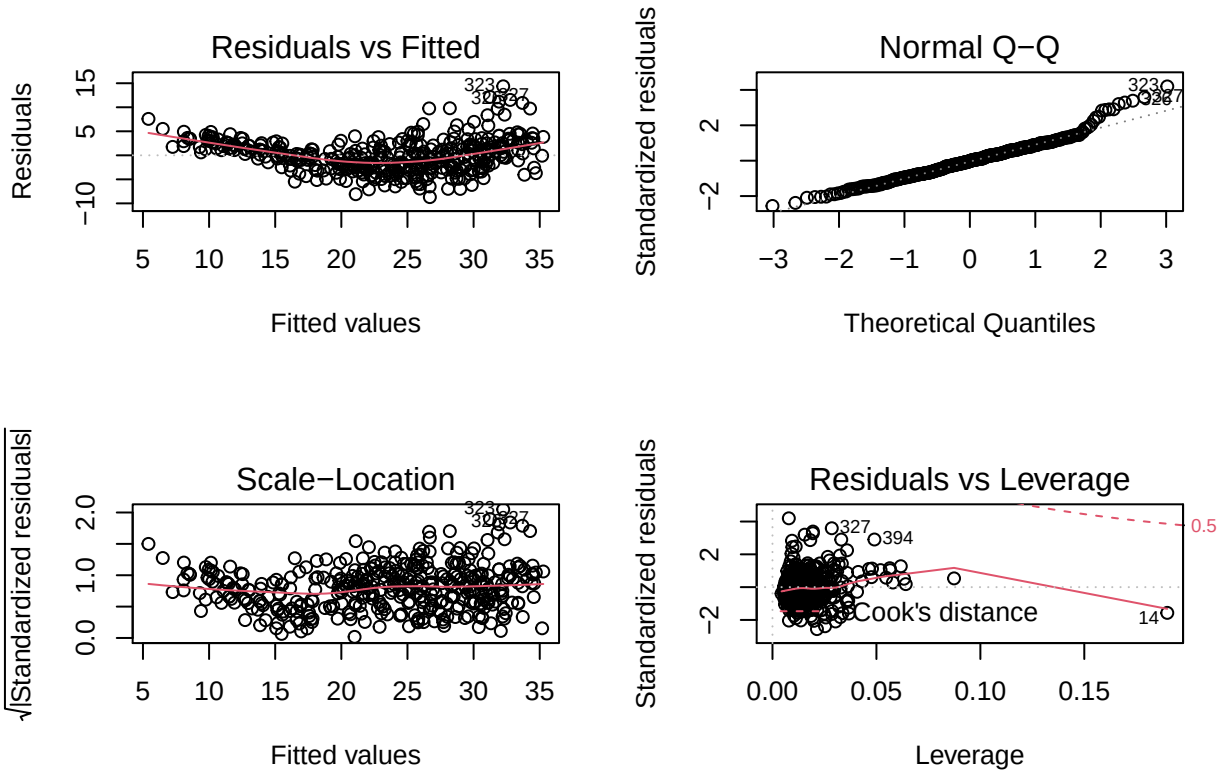
The most significant predictors appear to be: `weight` and `year` (each with p-value less than 0.001).

The coefficient of 0.75 on `year` suggests that for every one unit increase in model year there is a corresponding 0.75 unit increase `mpg`.

(d)

- The Residual plot suggests some evidence of non-linearity for the smallest and largest fitted values. Additionally, it shows a cluster of several outliers in the upper right corner of the plot.
- The QQ-plot suggests some significant deviation from Normality in the right tail of the distribution (in particular, the data has a heavier right tail than it would if Normally distributed)
- The scale-location plot suggests that variance of residuals is not constant, instead increasing as fitted values increase.
- The leverage plot indicates 1 relatively influential observation (#14). The outliers identified in the residual plot do not appear to be as influential.

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



(e) Of the (somewhat arbitrarily) selected interaction terms, all appear to be statistically significant.

```
mpg_mod2<-lm(mpg~. + year:weight + horsepower:acceleration + cylinders:displacement, data = Auto_q)

summary(mpg_mod2)
```

```
##
## Call:
## lm(formula = mpg ~ . + year:weight + horsepower:acceleration +
##     cylinders:displacement, data = Auto_q)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1925 -1.6451 -0.0395  1.3481 12.7188
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -9.047e+01  1.354e+01  -6.684 8.26e-11 ***
## cylinders               -1.394e+00  4.390e-01  -3.176 0.001614 **
## displacement            -7.384e-02  1.224e-02  -6.031 3.84e-09 ***
## horsepower               3.828e-02  2.604e-02   1.470 0.142421
## weight                   2.336e-02  4.493e-03   5.199 3.28e-07 ***
## acceleration             6.155e-01  1.590e-01   3.871 0.000127 ***
## year                     1.761e+00  1.691e-01  10.414  < 2e-16 ***
## weight:year             -3.659e-04  5.981e-05  -6.118 2.34e-09 ***
## horsepower:acceleration -6.721e-03  1.794e-03  -3.746 0.000207 ***
## cylinders:displacement   9.623e-03  1.727e-03   5.571 4.78e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.904 on 382 degrees of freedom
## Multiple R-squared:  0.8648, Adjusted R-squared:  0.8616
## F-statistic: 271.4 on 9 and 382 DF,  p-value: < 2.2e-16
```

(f) The pairwise scatterplots suggest that `displacement` and `horsepower` may have a non-linear relationship with `mpg`, so we apply log transformations to each. Doing so increased adjusted R^2 by about 0.02, and increased the significance of the `horsepower` predictor (P-vale < 0.001)

```
mpg_mod_transform<-lm(mpg ~ cylinders + I(log(displacement)) + I(log(horsepower))
                      + weight + acceleration + year, data = Auto_q
                                        )

summary(mpg_mod_transform)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + I(log(displacement)) + I(log(horsepower)) +
##     weight + acceleration + year, data = Auto_q)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1650  -1.8694  -0.1297   1.7799  13.0509
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          47.296757   9.325110   5.072 6.13e-07 ***
```

```
## cylinders              0.866307   0.291832   2.969 0.003180 **
## I(log(displacement)) -6.175953   1.174976  -5.256 2.44e-07 ***
## I(log(horsepower))   -8.008910   1.549491  -5.169 3.79e-07 ***
## weight               -0.002341   0.000686  -3.413 0.000711 ***
## acceleration         -0.391201   0.101825  -3.842 0.000143 ***
## year                  0.695901   0.048407  14.376  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.202 on 385 degrees of freedom
## Multiple R-squared:  0.8342, Adjusted R-squared:  0.8316
## F-statistic: 322.9 on 6 and 385 DF,  p-value: < 2.2e-16
```

Problem 6

Based on ISLR Exercise 3.14

This problem focuses on the *collinearity* problem.

(a) Run the following code, which randomly generates values for predictors for X_1 and X_2, and then generates values for a response Y based on a linear model. What are the regression coefficients for this model, as implied by the last line of code?

```
set.seed(1000)
n<- 100
x1 <- runif(n)
x2 <- 0.5*x1 + rnorm(n, mean = 0, sd = 0.1)
y  <- 2 + 2*x1+0.3*x2 + rnorm(n, mean = 0, sd = 1 )
```

(b) What is the correlation between `x1` and `x2`? Create a scatterplot showing this relationship.

(c) Using the simulated data, fit a least squares regression line for Y as a function of X_1 and X_2. What are the estimates for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$? How do these compare to the theoretical regression coefficients?

(d) Based on the regression summary, can you reject the null hypothesis $H_0 : \beta_1 = 0$? What about the null hypothesis $H_0 : \beta_2 = 0$? Explain.

(e) Now fit a least squares regression for Y as a function of just X_1. Based on this model, can you reject the null hypothesis $H_0 : \beta_1 = 0$?

(f) Similarly, fit a least squares regression for Y as a function of just X_2. Based on this model, can you reject the null hypothesis $H_0 : \beta_1 = 0$?

(g) Do the results in parts (d) - (f) contradict each other? Explain.

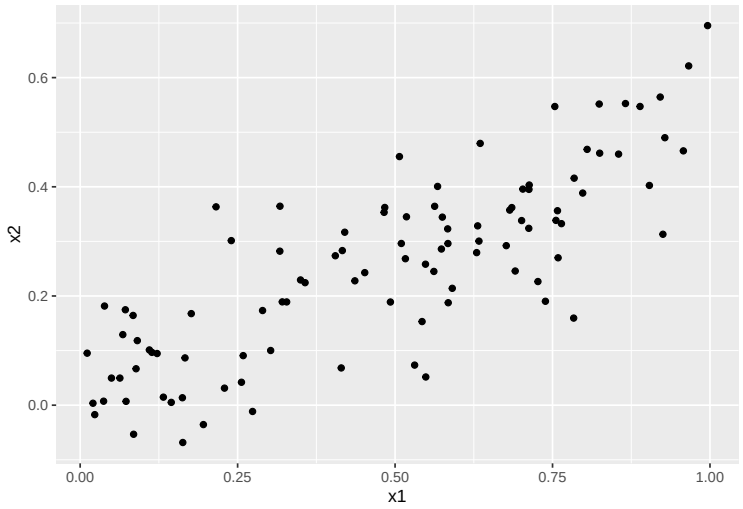
(a) Based on the model, the theoretical regression coefficients are $\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3$.

(b) The correlation between X_1 and X_2 is $r = 0.81$.

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

```
## [1] 0.8126056
```

```
prob6 <- data.frame(x1,x2,y)
ggplot(prob6, aes(x = x1, y= x2))+geom_point()
```



(c) The model estimates are $\hat{\beta}_0 = 1.69, \hat{\beta}_1 = 2.31, \hat{\beta}_2 = 0.81$, which are somewhat, but not extremely, close to the true model parameters.

```
p6_mod <- lm(y~x1+x2, data = prob6)
summary(p6_mod)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = prob6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29969 -0.66485  0.00941  0.64086  2.24372
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6925     0.1845   9.172 8.24e-15 ***
## x1            2.3173     0.5675   4.083 9.14e-05 ***
## x2            0.8165     0.9511   0.858    0.393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9294 on 97 degrees of freedom
## Multiple R-squared:  0.4122, Adjusted R-squared:  0.4001
## F-statistic: 34.01 on 2 and 97 DF,  p-value: 6.424e-12
```

(d) Based on the regression summary table, we can reject $H_0 : \beta_1 = 0$ since the p-value is less than 0.001, but cannot reject $H_0 : \beta_2 = 0$, since the p-value is relatively large.

(e) Based on the SLR model, we can reject $H_0 : \beta_1 = 0$.

```
p6_mod1 <- lm(y~x1, data = prob6)
summary(p6_mod1)
```

```
##
## Call:
## lm(formula = y ~ x1, data = prob6)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.31504 -0.62135 -0.02965  0.57487  2.14596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7090     0.1833   9.324 3.54e-15 ***
## x1            2.7132     0.3303   8.214 8.86e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9282 on 98 degrees of freedom
## Multiple R-squared:  0.4077, Adjusted R-squared:  0.4017
## F-statistic: 67.46 on 1 and 98 DF,  p-value: 8.86e-13
```

(f) Similarly, based on the SLR model, we can reject $H_0 : \beta_2 = 0$.

```
p6_mod2 <- lm(y~x2, data = prob6)
summary(p6_mod2)
```

```
##
## Call:
## lm(formula = y ~ x2, data = prob6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16544 -0.73182 -0.02886  0.67904  2.67635
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0053     0.1808  11.090  < 2e-16 ***
## x2            3.9723     0.5970   6.654 1.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.001 on 98 degrees of freedom
## Multiple R-squared:  0.3112, Adjusted R-squared:  0.3041
## F-statistic: 44.27 on 1 and 98 DF,  p-value: 1.649e-09
```

(g) The previous results are not contradictory, since the MLR model takes correlation of predictors into account. While individually, Y is correlated with both X_1 and X_2, most of variation in Y is explained by variation in X_1.
