# Homework 2

Rohan Krishnan

2024-01-29

## Problem 1

**No submission required for this problem.**

## Problem 2

### ISLR Chapter 3 Conceptual Exercise 3

(a). Answer iii is correct. We can deduce this answer by noting that $\beta_3$ is a strong positive integer (35), indicating that when an individual has gone to college (when $X_3$=1), their salary is predicted to increase by $35,000 holding all other factors constant. However, $\beta_5$, which is the interaction between GPA and level, is -10, meaning that if a college graduate as a GPA above 3.5, the marginal benefit of being college (35) is canceled out ($-10 * 1 * 3.5 = -35$) and begins to negatively impact their starting salary. Therefore, if their GPA is high enough (>3.5) and IQ and GPA are fixed, high school graduates do earn more on average.

(b). The predicted starting salary after graduation of a college graduate with IQ of 110 and GPA of 4.0 can be calculated by the given equation as follows:

$$\widehat{Salary} = 50 + (20 * 4.0) + (0.07 * 110) + (35 * 1) + (0.01 * 4.0 * 110) + (-10 * 4.0 * 1) = 137.1$$

As seen above, it is predicted that such an individual would earn $137,100 as their starting salary after graduation.

(c). **FALSE**. The only way to conclude that there is/is not evidence of an interaction effect is by examining the p-value produced by the regression's hypothesis test. If it is sufficiently small, we can conclude that there is statistically significant evidence of an interaction effect (even if the magnitude of the effect is small). It is also important to note that GPA and IQ are measured on very different scales, with GPAs falling between 1.0 and 4.0 and most IQs falling between 90 and 130.

### ISLR Chapter 4 Conceptual Exercise 4

(a). Since the cubic model contains the simple linear regression, it's training RSS would be at least as low as the simple model and with the addition of the quadratic and cubic terms will be lower, regardless of if the terms are useful or not.

(b). Without being able to see exploratory plots of the predictor's relationship to the response, there is not enough information to tell whether the testing RSS would be lower or higher. However, if there is not a cubic relationship, it is possible that the cubic model is over-fitted to the training data and will have a higher testing RSS.

(c). Knowing the relationship is not linear, we would still expect the cubic model to have a lower training RSS since it explains as much as the simple model plus the extra variation of the quadratic and cubic term.

(d). There is not enough information to tell if the testing RSS would be lower for the simple or cubic model since we do not know the exact type of non-linear relationship.

# Problem 3

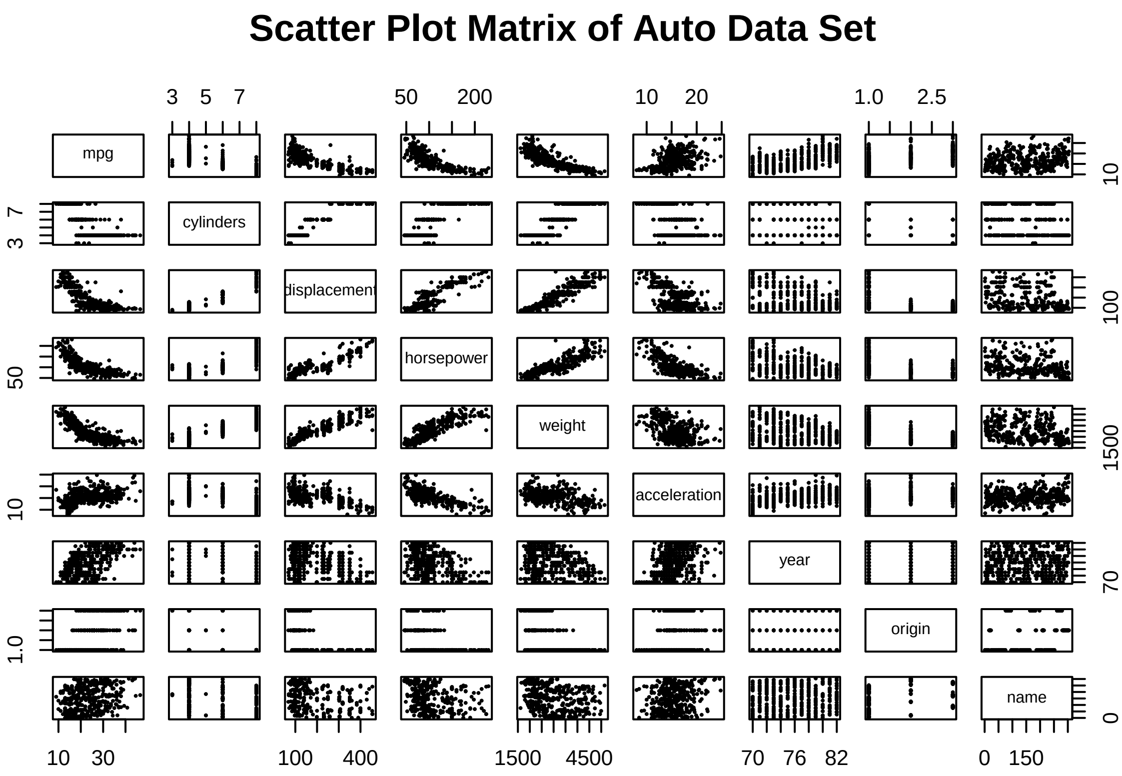
## ISLR Chapter 3 Applied Exercise 9

### Part A.

```
#Set working directory, load in Auto, remove any NA values
library(tidyverse)
```

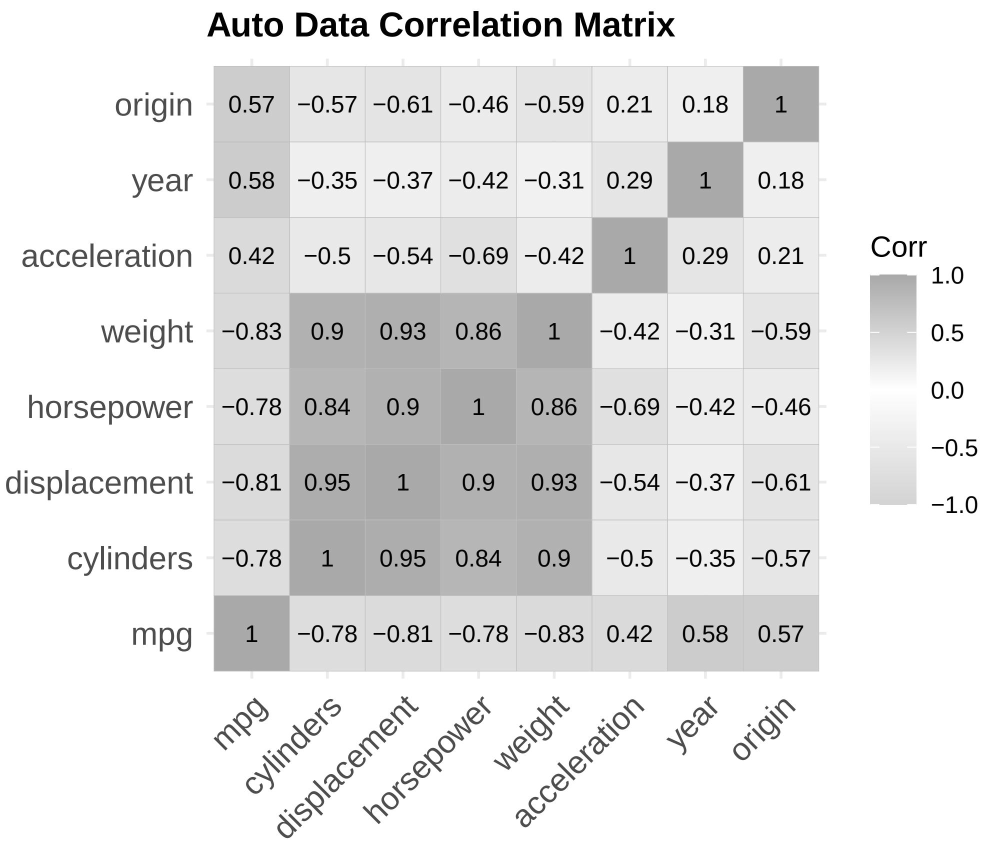
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
setwd("~/Downloads/")
Auto <- read.table("Auto.data", header = T, na.strings = "?", stringsAsFactors = T)
Auto <- na.omit(Auto)

#Generate scatter plot matrix of all variables in Auto
pairs(Auto, main = "Scatter Plot Matrix of Auto Data Set", cex = 0.2)
```



**Part B.**

```r
#Compute correlation matrix of all variables except "name"
correlations <- Auto %>%
  select(-name) %>%
  cor() %>%
  as.data.frame()

#Visualize correlations
library(ggcorrplot)

correlations %>%
  ggcorrplot(title = "Auto Data Correlation Matrix",
             colors = c("light grey", "white", "dark grey"),
             lab = TRUE, lab_col = "black", lab_size = 3) +
  theme(plot.title = element_text(face = "bold"))
```



**Part C.**

```r
#Fit multiple linear regression on mpg to all variables except name
lm.c <- lm(mpg ~ . - name, data = Auto)

#Create summary table of lm.c
library(stargazer, warn.conflicts = FALSE)
```

```
##
## Please cite as:
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
stargazer(lm.c, type = "latex", title = "Part C Multiple Linear Regression",
          header = FALSE, no.space = TRUE)
```

Table 1: Part C Multiple Linear Regression

| | *Dependent variable:* |
|---|---|
| | mpg |
| cylinders | −0.493 |
| | (0.323) |
| displacement | 0.020*** |
| | (0.008) |
| horsepower | −0.017 |
| | (0.014) |
| weight | −0.006*** |
| | (0.001) |
| acceleration | 0.081 |
| | (0.099) |
| year | 0.751*** |
| | (0.051) |
| origin | 1.426*** |
| | (0.278) |
| Constant | −17.218*** |
| | (4.644) |
| Observations | 392 |
| $R^2$ | 0.821 |
| Adjusted $R^2$ | 0.818 |
| Residual Std. Error | 3.328 (df = 384) |
| F Statistic | 252.428*** (df = 7; 384) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

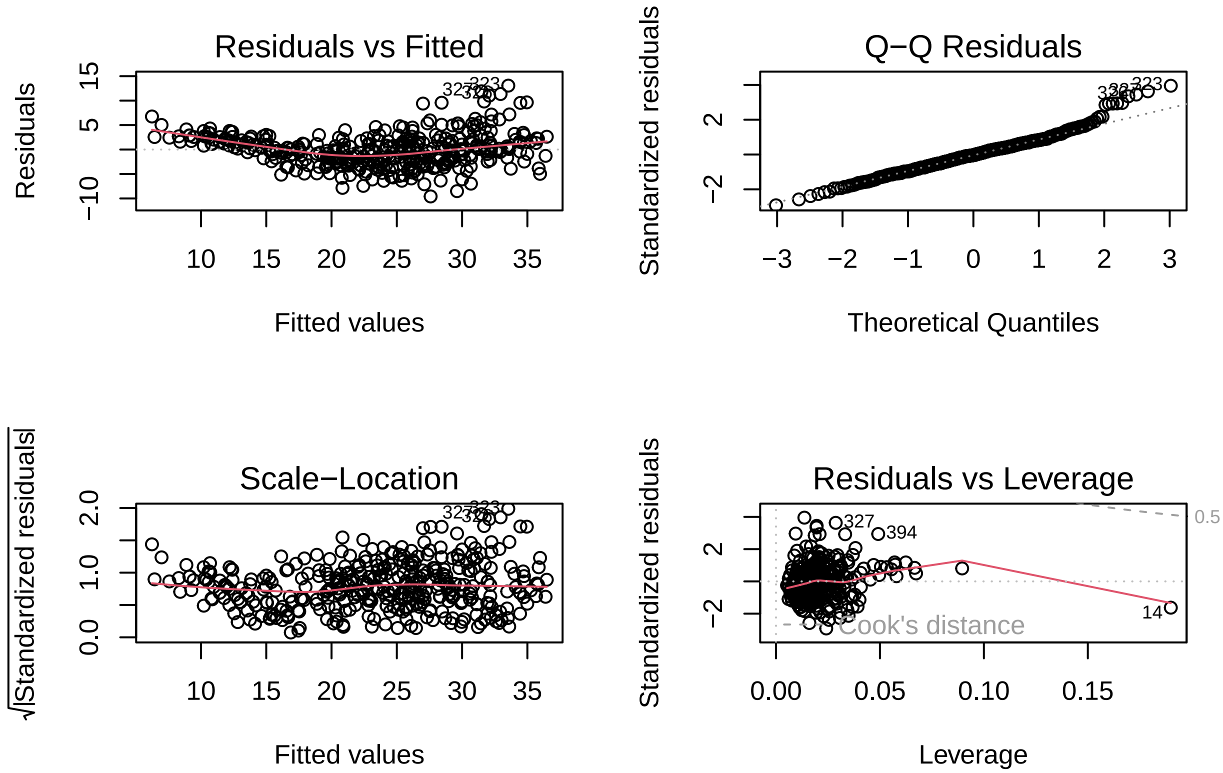
**(i).** There are negative relationships between mpg and cylinders, horsepower, and weight. There are positive relationship between mpg and displacement, acceleration, year, and origin.

**(ii).** Weight, year, and origin appear to have statistically significant relationships with mpg at the 99.9% confidence level and displacement seems to have a statistically significant relationship with mpg at the 95% confidence level.

**(iii)** Each increase in the model year of a car predicts a statistically significant 0.750773 increase in the car's miles per gallon holding all other features constant.

**Part D.**

```r
#Generate diagnostic plots of the above linear regression
par(mfrow = c(2,2))
plot(lm.c)
```




The residuals appear to be slightly heteroskedastic as there is a small trumpeting outwards as the fitted values increase. This could also be some type of trend, meaning that there is some relationship we missed with our regression. There appears to be a point (14) with unusually high leverage as well. There also appear to be a few larger outliers (residuals hitting ~10) among the higher fitted values.

**Part E.**

```r
#Create interaction term for cylinders and horsepower using * and view coefficient and p-value
lm.e1 <- lm(mpg ~. + cylinders*horsepower - name, data = Auto)

#Create interaction term for weight and acceleration using * and view coefficient and p-value
lm.e2 <- lm(mpg ~. + weight*acceleration - name, data = Auto)

#Create interaction term for year and origin using : and view coefficient and p-value
lm.e3 <- lm(mpg ~. + year:origin - name, data = Auto)

#Create summary table for above regressions
stargazer(lm.e1, lm.e2, lm.e3, type = "latex",
          title = "Part E MLR Interaction Effects", summary = FALSE,
          header = FALSE, no.space = TRUE)
```

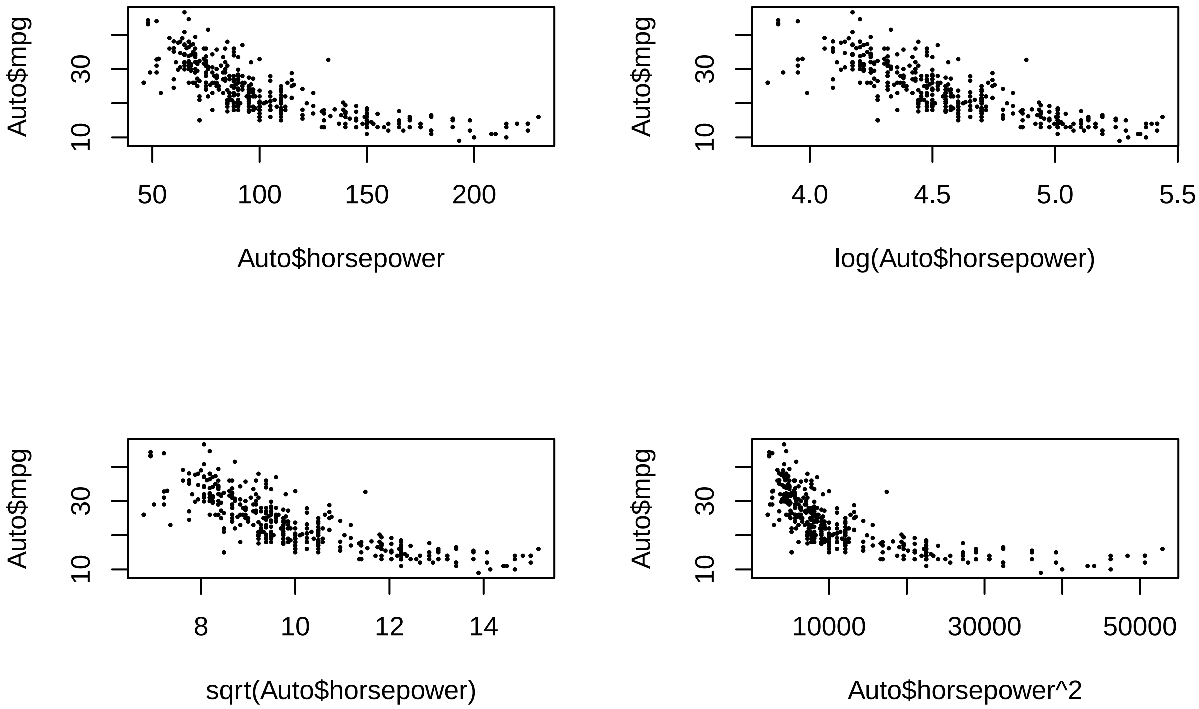
Table 2: Part E MLR Interaction Effects

| | *Dependent variable:* | | |
|---|---|---|---|
| | mpg | | |
| | (1) | (2) | (3) |
| cylinders | $-4.306^{***}$ | $-0.214$ | $-0.504$ |
| | (0.458) | (0.308) | (0.319) |
| displacement | $-0.001$ | 0.003 | $0.016^{**}$ |
| | (0.007) | (0.007) | (0.008) |
| horsepower | $-0.316^{***}$ | $-0.041^{***}$ | $-0.014$ |
| | (0.031) | (0.013) | (0.014) |
| weight | $-0.004^{***}$ | $0.004^{**}$ | $-0.006^{***}$ |
| | (0.001) | (0.002) | (0.001) |
| acceleration | $-0.170^{*}$ | $1.629^{***}$ | 0.092 |
| | (0.090) | (0.242) | (0.098) |
| year | $0.739^{***}$ | $0.782^{***}$ | $0.419^{***}$ |
| | (0.045) | (0.048) | (0.113) |
| origin | $0.903^{***}$ | $1.033^{***}$ | $-14.046^{***}$ |
| | (0.250) | (0.269) | (4.699) |
| cylinders:horsepower | $0.040^{***}$ | | |
| | (0.004) | | |
| weight:acceleration | | $-0.001^{***}$ | |
| | | (0.0001) | |
| year:origin | | | $0.199^{***}$ |
| | | | (0.060) |
| Constant | $11.703^{**}$ | $-43.641^{***}$ | 8.492 |
| | (4.912) | (5.811) | (9.044) |
| Observations | 392 | 392 | 392 |
| $R^2$ | 0.862 | 0.841 | 0.826 |
| Adjusted $R^2$ | 0.859 | 0.838 | 0.823 |
| Residual Std. Error (df = 383) | 2.929 | 3.141 | 3.286 |
| F Statistic (df = 8; 383) | $299.262^{***}$ | $253.908^{***}$ | $227.917^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | | |

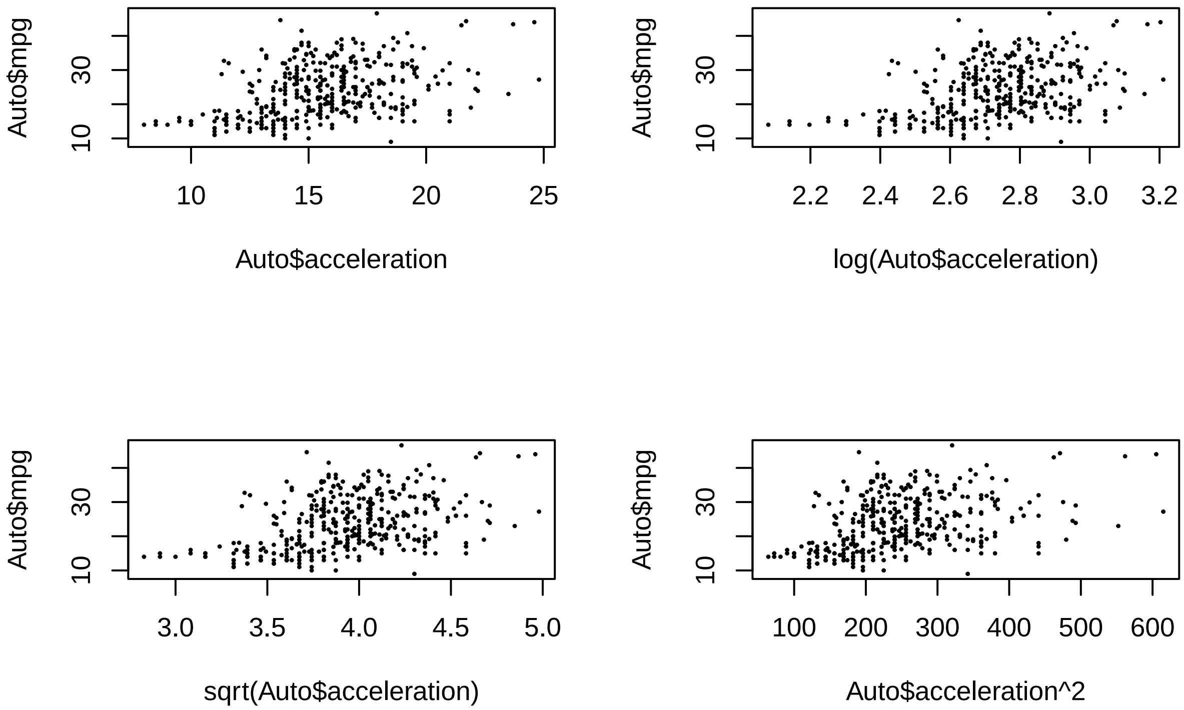
All of the interactions tested appear to be statistically significant at at least the 0.04 level. There are at least 3 significant cases of interactions between variables.

**Part F.**

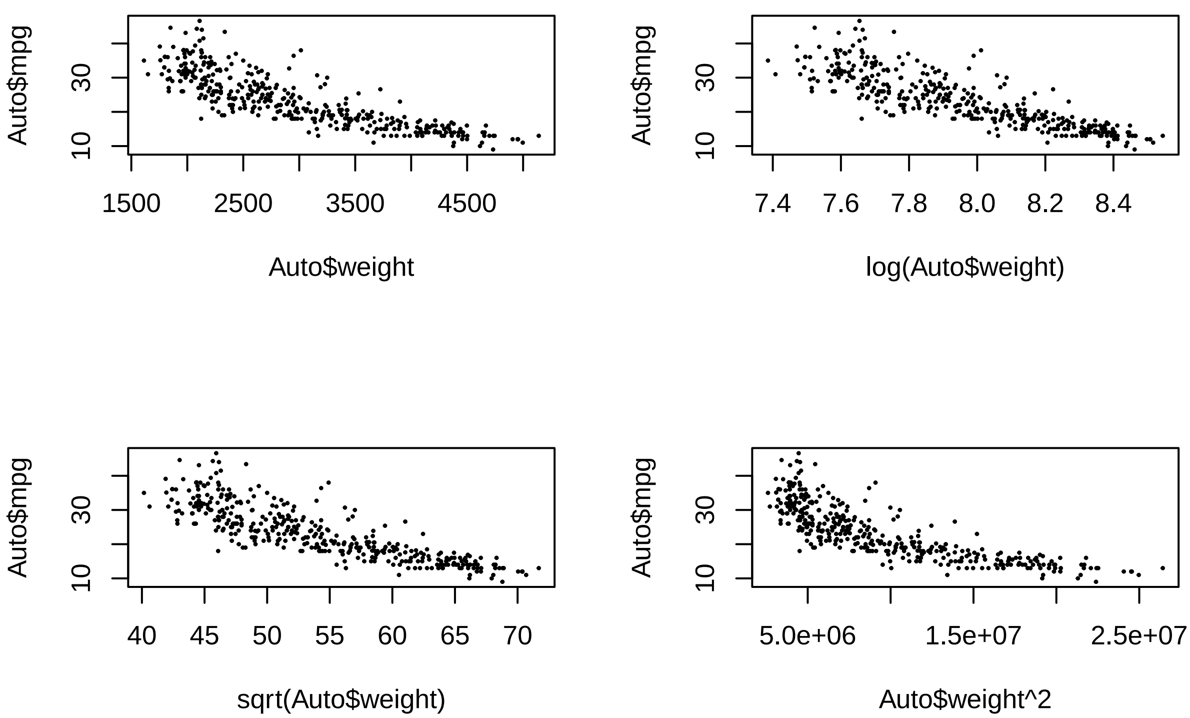
```r
#Log transform appears to give a more linear relationship
par(mfrow = c(2,2))
plot(Auto$horsepower, Auto$mpg, cex = 0.2)
plot(log(Auto$horsepower), Auto$mpg, cex = 0.2)
plot(sqrt(Auto$horsepower), Auto$mpg, cex = 0.2)
plot(Auto$horsepower^2, Auto$mpg, cex = 0.2)
```




```r
#None of the transformations appear to make the relationship more linear
par(mfrow = c(2,2))
plot(Auto$acceleration, Auto$mpg, cex = 0.2)
plot(log(Auto$acceleration), Auto$mpg, cex = 0.2)
plot(sqrt(Auto$acceleration), Auto$mpg, cex = 0.2)
plot(Auto$acceleration^2, Auto$mpg, cex = 0.2)
```




```r
#Again, log transform appears to make the relationship more linear
par(mfrow = c(2,2))
plot(Auto$weight, Auto$mpg, cex = 0.2)
plot(log(Auto$weight), Auto$mpg, cex = 0.2)
plot(sqrt(Auto$weight), Auto$mpg, cex = 0.2)
plot(Auto$weight^2, Auto$mpg, cex = 0.2)
```




For horsepower, the log transformation appeared to make the relationship between horsepower and mpg more linear. None of the transformations appeared to improve the linearity of the relationship between acceleration and mpg. The log transformation also appeared to work in making the relationship between weight and mpg more linear.

**ISLR Chapter 3 Applied Exercise 10**

**Part A.**

```r
#Load library
library(ISLR2)
```

```
##
## Attaching package: 'ISLR2'

## The following object is masked _by_ '.GlobalEnv':
##
##     Auto
```

```r
#Fit MLR of Sales ~ Price + Urban + US
lm.10a <- lm(Sales ~ Price + Urban + US, data = Carseats)
```

**Part B**

```r
#Generate summary table of lm.10a
stargazer(lm.10a, type = "latex", title = "Problem 10 B MLR Summary",
          header = FALSE, no.space = TRUE)
```

Table 3: Problem 10 B MLR Summary

| | *Dependent variable:* |
|---|---|
| | Sales |
| Price | $-0.054^{***}$ |
| | (0.005) |
| UrbanYes | $-0.022$ |
| | (0.272) |
| USYes | $1.201^{***}$ |
| | (0.259) |
| Constant | $13.043^{***}$ |
| | (0.651) |
| Observations | 400 |
| $R^2$ | 0.239 |
| Adjusted $R^2$ | 0.234 |
| Residual Std. Error | 2.472 (df = 396) |
| F Statistic | $41.519^{***}$ (df = 3; 396) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

A $1 (assumed unit of price measurement is in single dollars) increase in the price a company charges for car seats at each site predicts a statistically significant 0.054459 (sales measured in thousands so 0.054459 corresponds to 54.459) decrease in unit sales at each location holding all other factors constant. A store being in an urban location is not a statistically significant predictor of a change in sales at each location. A store being in the US predicts a statistically significant 1.200573 (sales measured in thousands so 1.200573 corresponds to 1200.573) increase in unit sales at each location holding all other factors constant.

**Part C.** Below is the model written out in equation form:

$$\widehat{Sales} = 13.04 + (-0.54 \times Price) + (-0.022 \times Urban) + (1.20 \times US)$$

If Urban = Yes, you set the Urban variable to 1. If not, it is set to 0. Similarly if US = Yes, it is set to 1 and 0 otherwise. Below are all of the possible models based on Urban and US values.

$$\widehat{Sales} = 13.04 + (-0.54 \times Price) + \begin{cases} -0.022, & \text{if Urban is Yes and US is No} \\ 1.20, & \text{if Urban is No and US is Yes} \\ 1.18, & \text{if Urban is Yes and US is Yes} \\ 0, & \text{if Urban is No and US is No} \end{cases}$$

**Part D.**

Price and US appear to have $\hat{\beta}s$ statistically significantly different than 0. Urban has a large p-value indicating that there is not sufficient evidence to reject the null hypothesis.

**Part E.**

```
#Run regression of Sales on Price and US and generate summary table
lm.10d <- lm(Sales ~ Price + US, data = Carseats)
stargazer(lm.10d, type = "latex", title = "Problem 10 E MLR Summary",
          header = FALSE, no.space = TRUE)
```

Table 4: Problem 10 E MLR Summary

| | *Dependent variable:* |
|---|---|
| | Sales |
| Price | $-0.054^{***}$ |
| | (0.005) |
| USYes | $1.200^{***}$ |
| | (0.258) |
| Constant | $13.031^{***}$ |
| | (0.631) |
| Observations | 400 |
| $R^2$ | 0.239 |
| Adjusted $R^2$ | 0.235 |
| Residual Std. Error | 2.469 (df = 397) |
| F Statistic | $62.431^{***}$ (df = 2; 397) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Part F.**

```
#Compare two regressions and run anova to see if one is a better fit
stargazer(lm.10a, lm.10d, type = "latex", title = "Problem 10 F MLR Comparison",
          header = FALSE, no.space = TRUE)
```

```
anova(lm.10a, lm.10d)
```

Analysis of Variance Table

Model 1: Sales ~ Price + Urban + US Model 2: Sales ~ Price + US Res.Df RSS Df Sum of Sq F Pr(>F) 1 396 2420.8
2 397 2420.9 -1 -0.03979 0.0065 0.9357

The models have similar $R^2$ values and coefficient values for each Price and USYes. The anova test shows that the model containing Urban does not significantly outperform the model that does not contain Urban.

Table 5: Problem 10 F MLR Comparison

| | *Dependent variable:* | |
|---|---|---|
| | Sales | |
| | (1) | (2) |
| Price | $-0.054^{***}$ | $-0.054^{***}$ |
| | (0.005) | (0.005) |
| UrbanYes | $-0.022$ | |
| | (0.272) | |
| USYes | $1.201^{***}$ | $1.200^{***}$ |
| | (0.259) | (0.258) |
| Constant | $13.043^{***}$ | $13.031^{***}$ |
| | (0.651) | (0.631) |
| Observations | 400 | 400 |
| $R^2$ | 0.239 | 0.239 |
| Adjusted $R^2$ | 0.234 | 0.235 |
| Residual Std. Error | 2.472 (df = 396) | 2.469 (df = 397) |
| F Statistic | $41.519^{***}$ (df = 3; 396) | $62.431^{***}$ (df = 2; 397) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

The model with Urban will always have a higher SSR because each additional variable will explain some variance even if it is not significant.

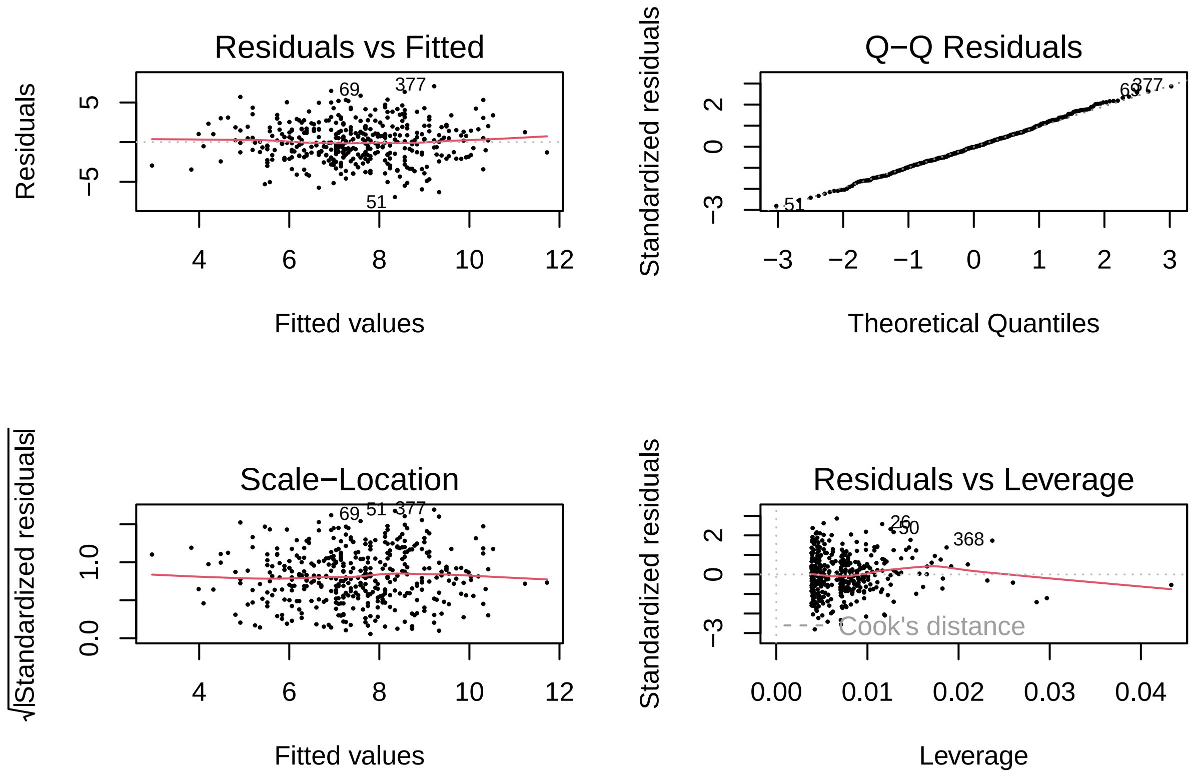
**Part G.**

```r
#Generate confidence interval
confint(lm.10d)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

**Part H.**

```r
#Plot regression from 10 D
par(mfrow = c(2,2))
plot(lm.10d, cex = 0.2)
```




There appear to be a couple high leverage observations in the model and one or two outliers.

### ISLR Chapter 3 Applied Exercise 13

#### Part A.

```r
#Set seed and create random normal variable x
set.seed(1)
x <- rnorm(100)
```

#### Part B.

```r
#Set seed and create random normal variable eps
set.seed(5)
eps <- rnorm(100, mean = 0, sd = 0.25)
```

#### Part C.

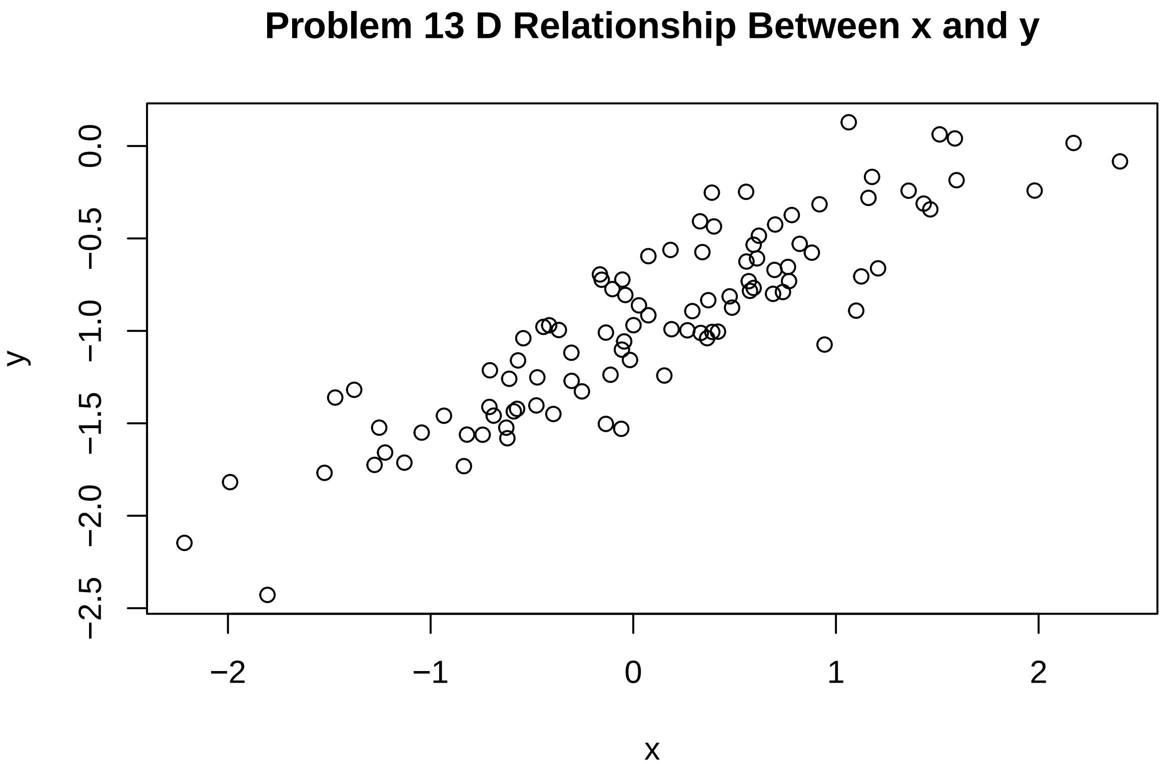
```r
#Create y as a linear function of x and eps and find length of y
y <- -1 + (0.5*x) + eps; length(y)
```

```
## [1] 100
```

The length of y is 100. $\beta_0$ is -1 and $\beta_1$ is 0.50.

#### Part D.

```r
#Plot x and y
plot(x = x, y = y, main = "Problem 13 D Relationship Between x and y")
```




The appears to be a fairly pronounced positive linear relationship between x and y.

**Part E.**

```r
#Generate regression of y on x and summarize
lm.13e <- lm(y ~x)
stargazer(lm.13e, type = "latex", no.space = TRUE,
          title = "Problem 13 E SLR Summary", header = FALSE)
```

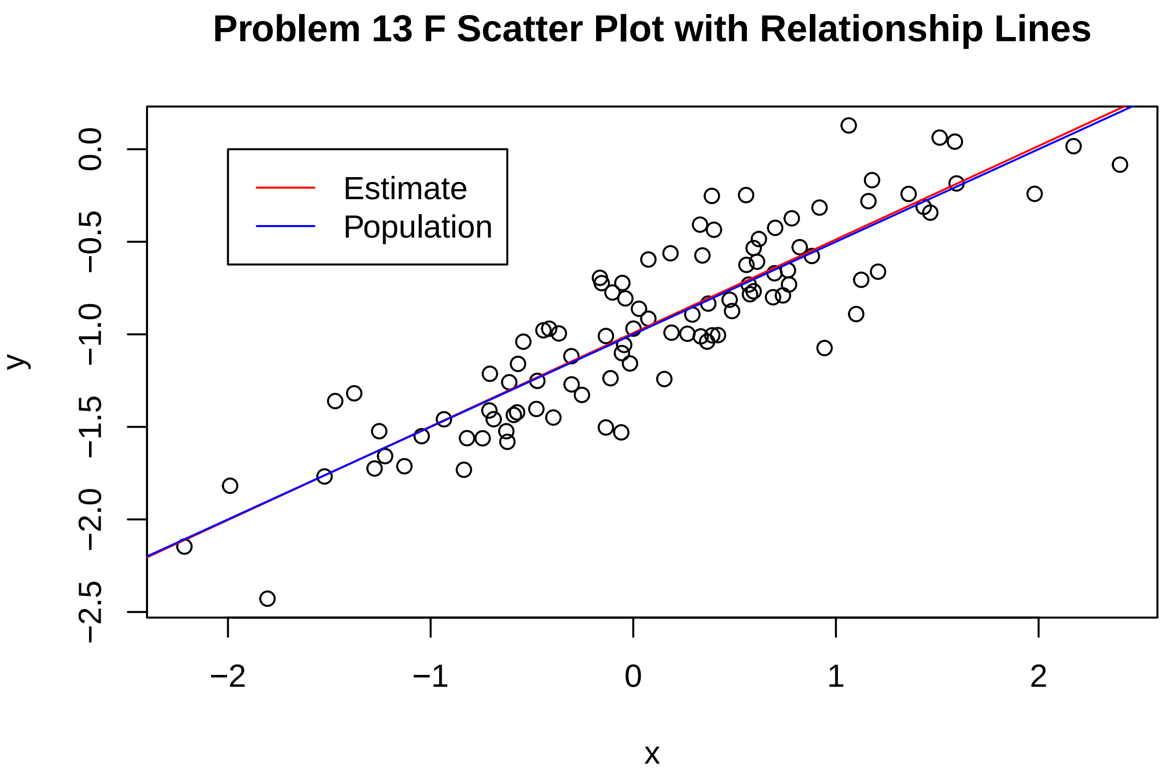
Table 6: Problem 13 E SLR Summary

| | *Dependent variable:* |
|---|---|
| | y |
| x | 0.505*** |
| | (0.027) |
| Constant | −0.993*** |
| | (0.024) |
| Observations | 100 |
| $R^2$ | 0.787 |
| Adjusted $R^2$ | 0.785 |
| Residual Std. Error | 0.237 (df = 98) |
| F Statistic | 361.403*** (df = 1; 98) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

$\hat{\beta}_0$ is -0.99 which is extremely close to $\beta_0$ (which is -1). $\hat{\beta}_1$ is 0.51 which also extremely close to $\beta_1$ (which is 0.50).

**Part F.**

```r
#Plot x vs y and compare true equation with regression
plot(x = x, y = y, main = "Problem 13 F Scatter Plot with Relationship Lines")
abline(lm.13e, col = "red")
abline(coef = c(-1,0.50), col = "blue")
legend(-2, 0, c("Estimate", "Population"), col = c("red", "blue"),lty = 1)
```




**Part G.**

```r
#Generate quadratic regression and compare with non-quadratic regression
lm.13g <- lm(y~x + I(x^2))
stargazer(lm.13e, lm.13g, type = "latex", no.space = TRUE,
          title = "Comparing 13 E and 13 G Regressions", header = FALSE)
```

There is not a statistically significant correlation between $x^2$ and y. We know this because the t-test of its coefficient yeilded a p-value of 0.299, which is too large to be considered statistically significant at any level.

**Part H.**

```r
#Sed seed and create new 'less noisy' eps and y
set.seed(5)
eps_less <- rnorm(100, mean = 0 , sd = 0.05)
y_less <- -1 + (0.5*x) +eps_less

#Generate new regression
lm.13h <- lm(y_less~x)

#Compare with normal noise data
stargazer(lm.13e, lm.13h, type = "latex", no.space = TRUE,
          title = "Comparing Regular vs Less Noise Models", header = FALSE)
```

```r
#Plot less noisy regression and compare true vs estimated relationship
plot(x, y_less, main = "Problem 13 I Model Comparison")
```

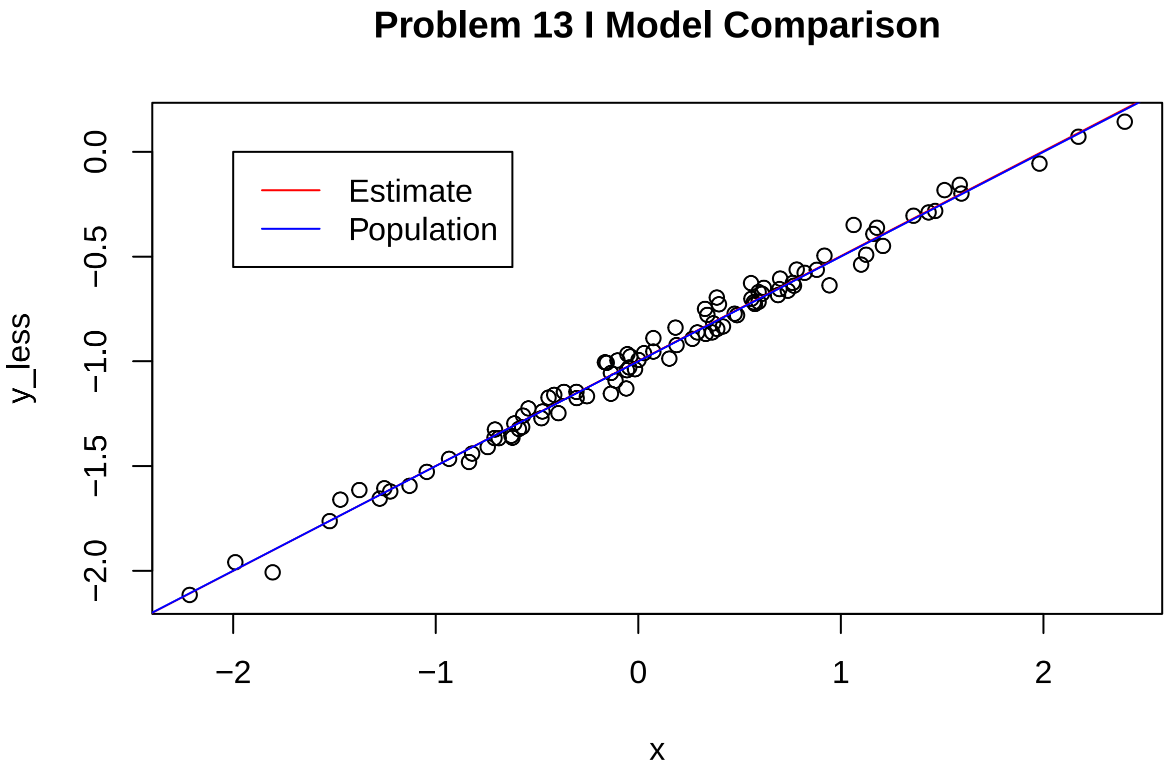
Table 7: Comparing 13 E and 13 G Regressions

| | *Dependent variable:* | |
|---|---|---|
| | y | |
| | (1) | (2) |
| x | 0.505*** | 0.509*** |
| | (0.027) | (0.027) |
| I(x^2) | | −0.022 |
| | | (0.021) |
| Constant | −0.993*** | −0.975*** |
| | (0.024) | (0.029) |
| Observations | 100 | 100 |
| $R^2$ | 0.787 | 0.789 |
| Adjusted $R^2$ | 0.785 | 0.785 |
| Residual Std. Error | 0.237 (df = 98) | 0.237 (df = 97) |
| F Statistic | 361.403*** (df = 1; 98) | 181.413*** (df = 2; 97) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

Table 8: Comparing Regular vs Less Noise Models

| | *Dependent variable:* | |
|---|---|---|
| | y | y_less |
| | (1) | (2) |
| x | 0.505*** | 0.501*** |
| | (0.027) | (0.005) |
| Constant | −0.993*** | −0.999*** |
| | (0.024) | (0.005) |
| Observations | 100 | 100 |
| $R^2$ | 0.787 | 0.989 |
| Adjusted $R^2$ | 0.785 | 0.989 |
| Residual Std. Error (df = 98) | 0.237 | 0.047 |
| F Statistic (df = 1; 98) | 361.403*** | 8,888.011*** |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

```r
abline(lm.13h, col = "red")
abline(coef = c(-1,0.5), col = "blue")
legend(-2, 0, c("Estimate", "Population"), col = c("red", "blue"),lty = 1)
```




The data has a tighter positive linear relationship. The model fitted on the less noisy data also has a much higher $R^2$ value.

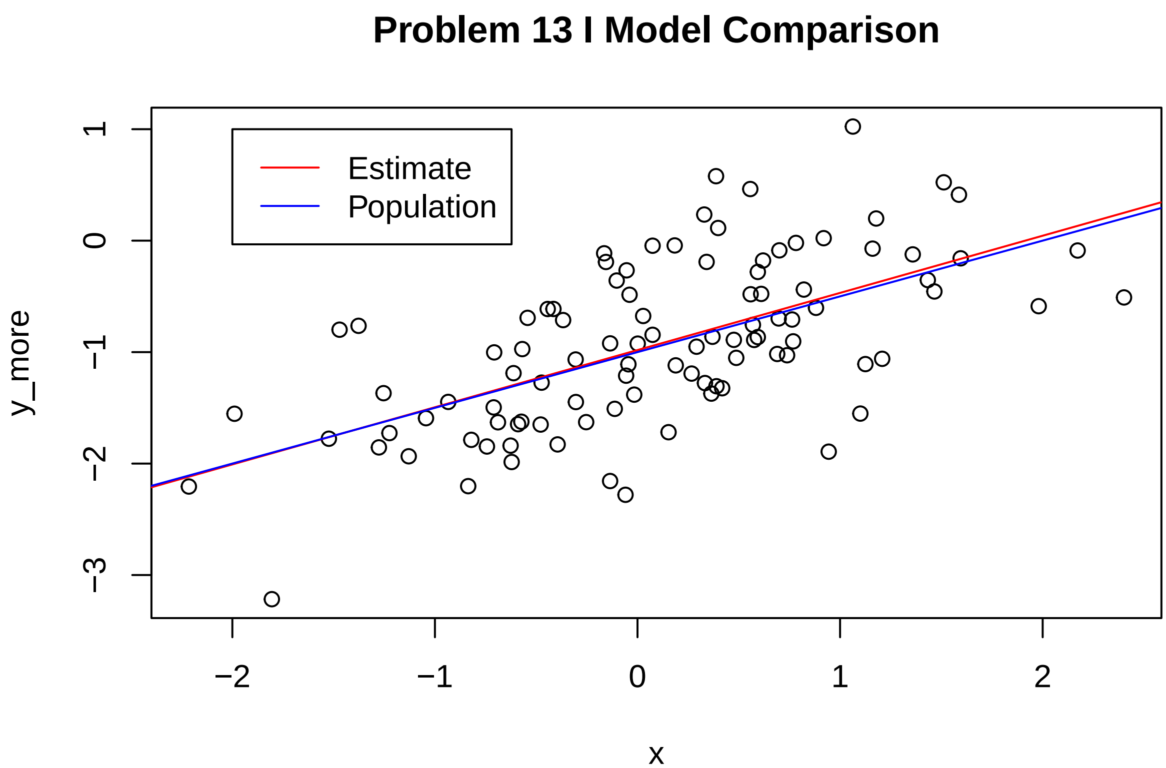
**Part I.**

```r
#Set seed and create new 'more noisy' eps and y
set.seed(5)
eps_more <- rnorm(100, mean = 0, sd = 0.625)
y_more <- -1 + (0.50*x) + eps_more

#Generate new regression
lm.13i <- lm(y_more ~ x)

#Compare normal, less, and more noisy regressions
stargazer(lm.13e, lm.13h, lm.13i, type = "latex", no.space = TRUE,
          title = "Comparing Regular vs Less vs More Noise", header = FALSE,
          omit.stat = "all")
```

```r
#Plot more noisy data and compare true and estimated relationship
plot(x, y_more, main = "Problem 13 I Model Comparison")
abline(lm.13i, col = "red")
abline(coef = c(-1,0.5), col = "blue")
legend(-2, 1, c("Estimate", "Population"), col = c("red", "blue"),lty = 1)
```

Table 9: Comparing Regular vs Less vs More Noise

| | Dependent variable: | | |
|---|---|---|---|
| | y | y_less | y_more |
| | (1) | (2) | (3) |
| x | 0.505*** | 0.501*** | 0.513*** |
| | (0.027) | (0.005) | (0.066) |
| Constant | −0.993*** | −0.999*** | −0.982*** |
| | (0.024) | (0.005) | (0.060) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |




The data shows a much weaker positive linear relationship between x and y. The model fitted on the noiser data also has a lower $R^2$ value compared to the other two models.

**Part J.**

```r
#Generate confidence intervals
confint(lm.13e)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0401284 -0.9451779
## x            0.4524277  0.5578924
```

```r
confint(lm.13h)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0080257 -0.9890356
## x            0.4904855  0.5115785
```

```r
confint(lm.13i)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1003209 -0.8629446
## x            0.3810693  0.6447311
```

The confidence intervals narrow for the model fitted on less noisy data and widen for the model fitted on more noisy data.

### ISLR Chapter 3 Applied Exercise 14

#### Part A.

```r
#Set seed and generate random variables
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100) / 10
y <- 2 + 2 * x1 + 0.3 * x2 + rnorm(100)
```

The form of the linear model is as follows:

$$y = 2 + 2 \times x_1 + 0.3 \times x_2 + \epsilon$$

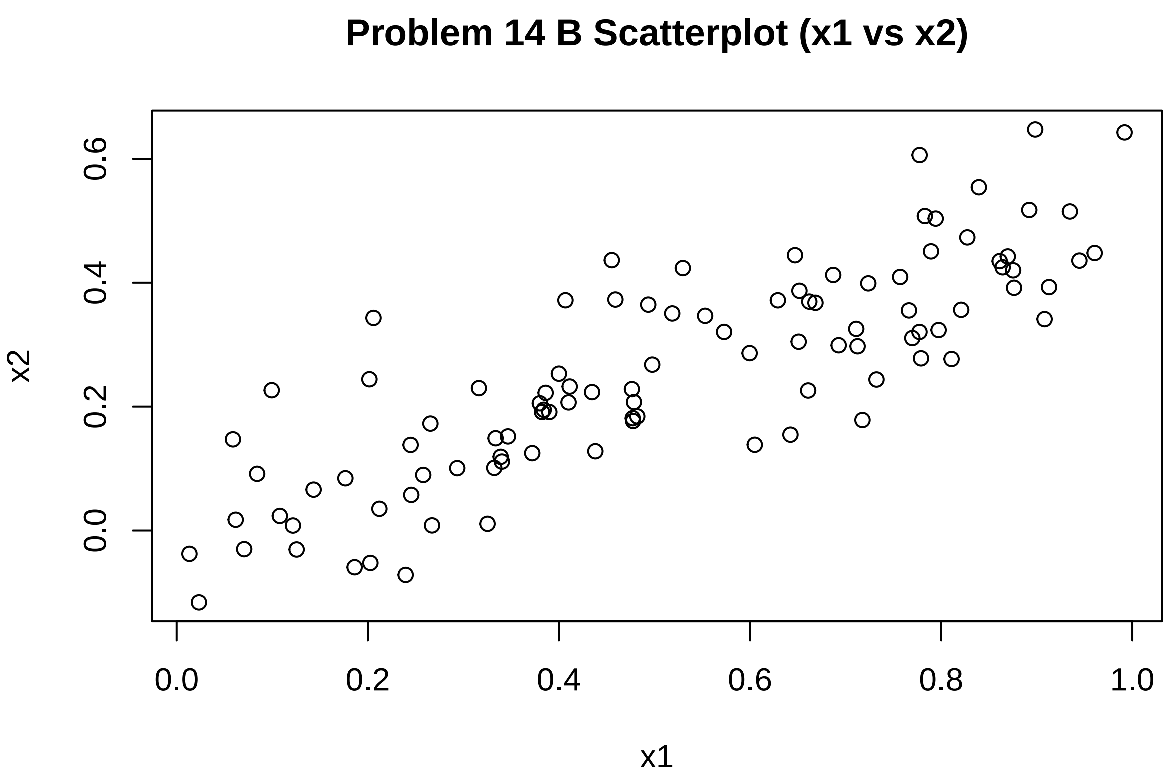
$\beta_0$ is 2, $\beta_1$ is 2, and $\beta_2$ is 0.3.

#### Part B.

```r
#Find correlation between x1 and x2 and plot
cor(x1, x2)
```

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

```r
plot(x1, x2, main = "Problem 14 B Scatterplot (x1 vs x2)")
```



From Part A, we know that $x_2$ is a multiple of $x_1$ with some noise added on. In other words, $x_2$ is a linear function of $x_1$. The correlation between $x_1$ and $x_2$ is 0.835. As is shown in the scatter plot, $x_1$ and $x_2$ have a positive correlation.

**Part C.**

```
#Generate colinear regression and summarize
lm.14c <- lm(y ~ x1 + x2)

stargazer(lm.14c, type = "latex", no.space = TRUE,
          title = "Problem 14 C Regression Summary", header = FALSE,
          omit.stat = "all")
```

Table 10: Problem 14 C Regression Summary

| | *Dependent variable:* |
|---|---|
| | y |
| x1 | 1.440** |
| | (0.721) |
| x2 | 1.010 |
| | (1.134) |
| Constant | 2.130*** |
| | (0.232) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

$\hat{\beta}_0$ is 2.13, $\hat{\beta}_1$ is 1.44, and $\hat{\beta}_2$ is 1.01. $\hat{\beta}_0$ (2.13) is quite close to $\beta_0$ (2). $\hat{\beta}_1$ (1.44) is somewhat close to $\beta_1$ (2) and $\hat{\beta}_2$ (1.01) is not very close to $\beta_2$ (0.30). From the summary table, we can reject the null hypothesis that $beta_1$ equals 0 only at the 0.05 level. In other words, we can say that $\beta_1$ is statistically significantly different from 0 with 95% confidence. However, we cannot reject the null hypothesis that $\beta_2$ equals 0, as it has a p-value of 0.3754 which is well above our minimum threshold of 0.05.

**Part D.**

```
#Generate SLR and summarize
lm.14d <- lm(y ~ x1)

stargazer(lm.14c, lm.14d, type = "latex", no.space = TRUE,
          title = "Comparing Colinear Regression with SLR", header = FALSE,
          omit.stat = "all")
```

The coefficient for $x_1$ shoots up to 1.98 which is much closer to its true value of 2. We can now reject the null hypothesis that $\beta_1$ equals 0 at the 0.001 level.

**Part E.**

```
#Generate other SLR and summarize
lm.14e <- lm(y ~ x2)

stargazer(lm.14c, lm.14d, lm.14e, type = "latex", no.space = TRUE,
          title = "Comparing Colinear Regression with SLRs", header = FALSE,
          omit.stat = "all")
```

The estimate for $x_2$ shoots up to 2.90 which is significantly higher than the model with both variables (1.01) but even farther away from its true value of 0.30. However, in this regression, we can reject the null hypothesis that $\beta_2$ equals 0 at the 0.001 level.

Table 11: Comparing Colinear Regression with SLR

| | *Dependent variable:* | |
|---|---|---|
| | y | |
| | (1) | (2) |
| x1 | $1.440^{**}$ | $1.976^{***}$ |
| | (0.721) | (0.396) |
| x2 | 1.010 | |
| | (1.134) | |
| Constant | $2.130^{***}$ | $2.112^{***}$ |
| | (0.232) | (0.231) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 12: Comparing Colinear Regression with SLRs

| | *Dependent variable:* | | |
|---|---|---|---|
| | y | | |
| | (1) | (2) | (3) |
| x1 | $1.440^{**}$ | $1.976^{***}$ | |
| | (0.721) | (0.396) | |
| x2 | 1.010 | | $2.900^{***}$ |
| | (1.134) | | (0.633) |
| Constant | $2.130^{***}$ | $2.112^{***}$ | $2.390^{***}$ |
| | (0.232) | (0.231) | (0.195) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | | |

**Part F.**

No because both variables are strongly correlated. They are each individually able to predict a large amount of the variability in y. However, since $x_2$ is a linear transformation of $x_1$, it is hard to parse out what explainability comes from what variable when both are used in the regression together.

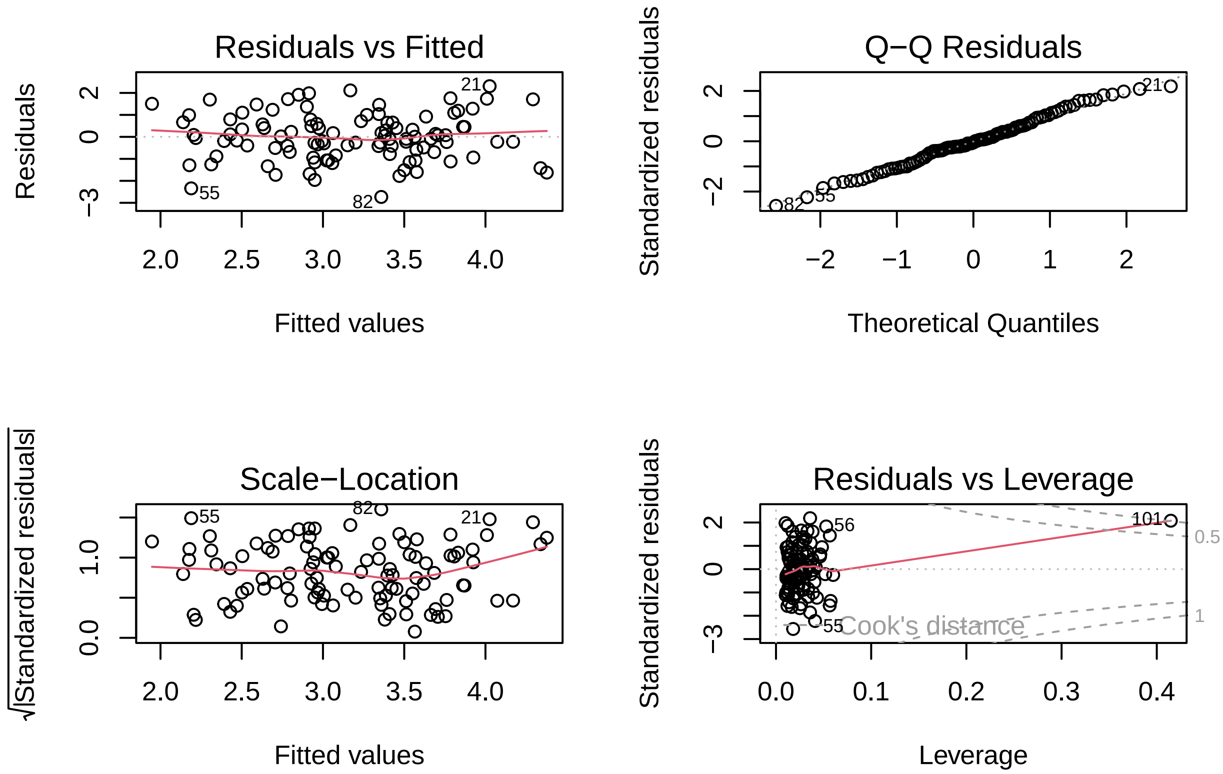
**Part G.**

```r
#Update vectors with mismeasured observations
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)

#Generate new regressions
lm.14g1 <- lm(y ~ x1 + x2)
lm.14g2 <- lm(y ~ x1)
lm.14g3 <- lm(y ~ x2)

#Compare regressions
stargazer(lm.14g1, lm.14g2, lm.14g3, type = "latex", no.space = TRUE,
          title = "Mismeasured Observation Regression Comparison", header = FALSE)
```

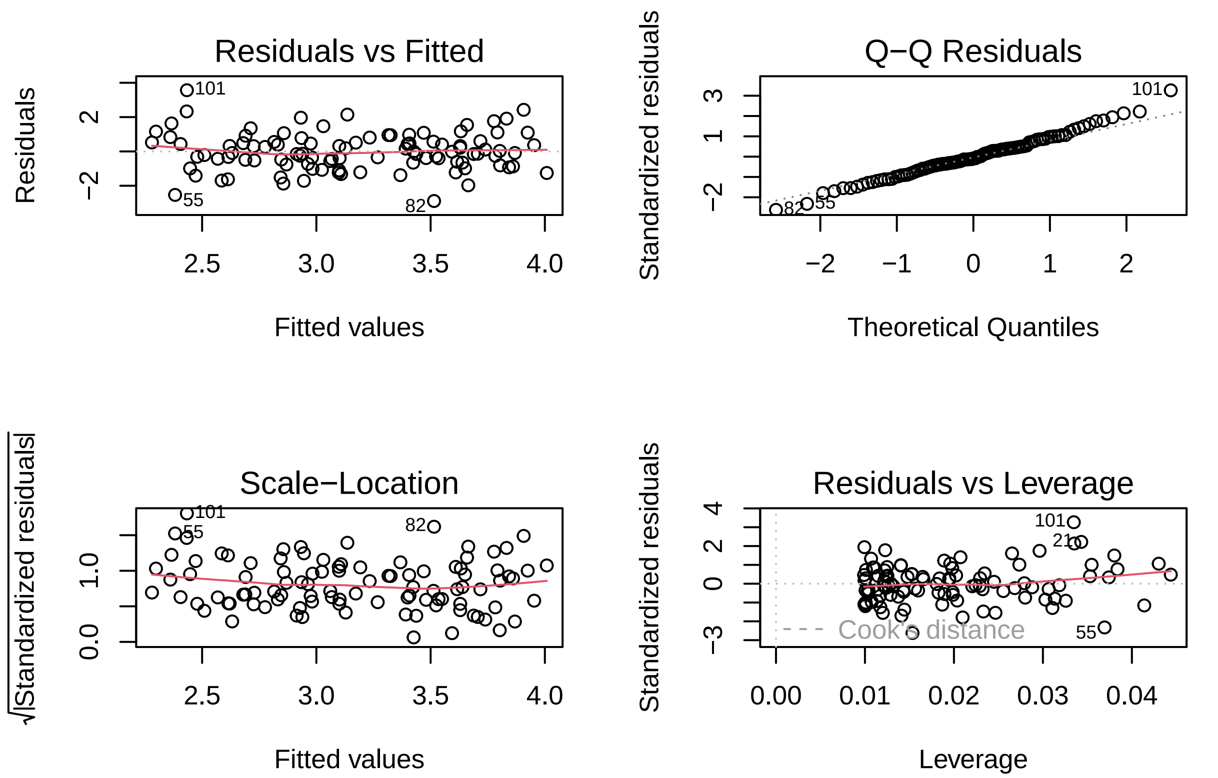
Table 13: Mismeasured Observation Regression Comparison

| | *Dependent variable:* | | |
|---|---|---|---|
| | y | | |
| | (1) | (2) | (3) |
| x1 | 0.539 | 1.766*** | |
| | (0.592) | (0.412) | |
| x2 | 2.515*** | | 3.119*** |
| | (0.898) | | (0.604) |
| Constant | 2.227*** | 2.257*** | 2.345*** |
| | (0.231) | (0.239) | (0.191) |
| Observations | 101 | 101 | 101 |
| $R^2$ | 0.219 | 0.156 | 0.212 |
| Adjusted $R^2$ | 0.203 | 0.148 | 0.204 |
| Residual Std. Error | 1.075 (df = 98) | 1.111 (df = 99) | 1.074 (df = 99) |
| F Statistic | 13.724*** (df = 2; 98) | 18.333*** (df = 1; 99) | 26.664*** (df = 1; 99) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

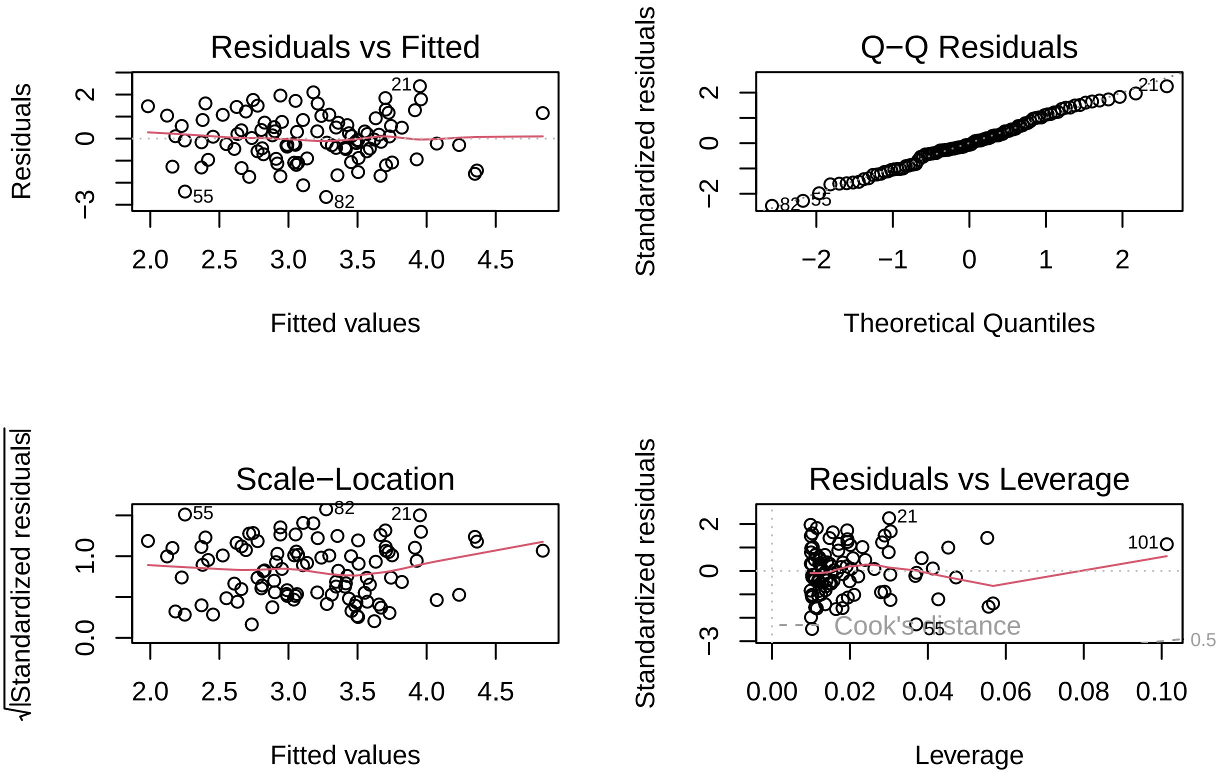
```r
#Plot each regression
par(mfrow = c(2,2))
plot(lm.14g1)
```




```
par(mfrow = c(2,2))
plot(lm.14g2)
```




```
par(mfrow = c(2,2))
plot(lm.14g3)
```




In the model with both predictors, the point has very high leverage but is not an outlier. It also drastically reduced $\hat{\beta}_1$ and drastically increased $\hat{\beta}_2$ while making $\hat{\beta}_1$ not statistically significant from 0. In the model with just $x_1$, it is an outlier but does not have high leverage. It also slightly reduced $\hat{\beta}_1$. Finally, in the model with only $x_2$, it has high leverage and is not an outlier. It also greatly increased $\hat{\beta}_2$ and made it statistically significantly different from 0. All of the models had lower $R^2$ values.

## Problem 4

### Part A.

```r
#Set seed so random samples are reproducible
set.seed(100)

#Create 25x25 matrix of random standard normal values and convert to df
df.train <- matrix(rnorm(625), nrow = 25)
df.train <- data.frame(df.train)
colnames(df.train)[1] <- "y"
```

### Part B.

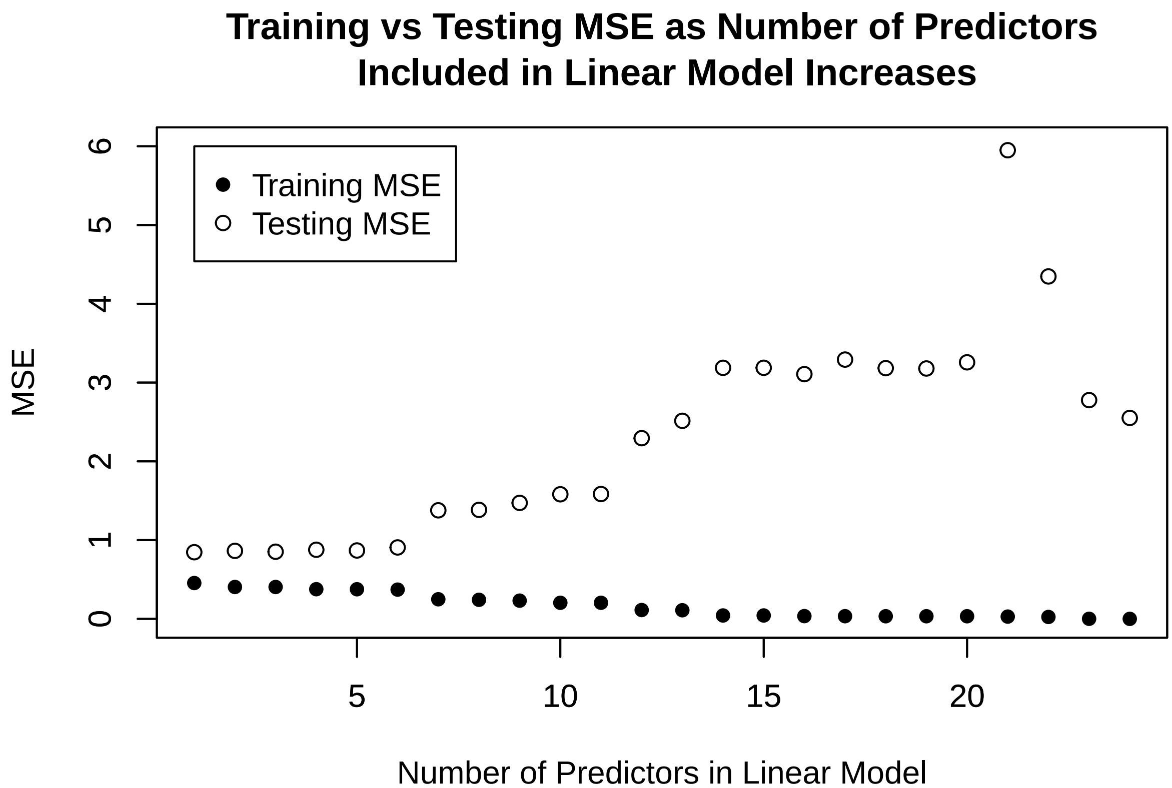
```r
#Set seed so random samples are reproducible
set.seed(50)

#Create 25x25 matrix of random standard normal values and convert to df
df.test <- matrix(rnorm(625), nrow = 25)
df.test <- data.frame(df.test)
colnames(df.test)[1] <- "y"
```

**Part C.**

```r
#Initalize MSE.train and MSE.test vectors
MSE.train <- vector()
MSE.test <- vector()

#Write function to iteratively linearly regress on y while adding an additional predictor
for (i in 2:ncol(df.train)) {
  #Create interim data frame with only columns corresponding to i
  interim.df <- df.train[,c(1:i)]

  #Fit model
  lm.fit <- lm(y ~ ., data = interim.df)

  #Calculate training MSE
  MSE.train[i-1] <- mean(lm.fit$residuals^2)

  #Create evaluation data frame of residuals of predicted values from df.test
  eval.df <- data.frame(residuals =  df.test$y - predict(lm.fit, df.test))

  #Calculate testing MSE
  MSE.test[i-1] <- mean((eval.df$residuals)^2)
}
```

**Part D.**

```r
#Generate training MSE plot
plot(x = 1:24  , y = MSE.train, col = "black",
     pch = 16, xlab = "Number of Predictors in Linear Model",
     ylab = "MSE" , ylim = c(0,6),
     main = "Training vs Testing MSE as Number of Predictors\n Included in Linear Model Increases")

#Overlay testing MSE on same plot
par(new = TRUE)
plot(x = 1:24, y = MSE.test, col = "black", xlab = "", ylab = "", ylim = c(0,6))

#Add legend
legend(1, 6, legend = c("Training MSE", "Testing MSE"),
       col = c("black", "black"), pch = c(16, 1), cex = 1)
```




**Part E.**

As the number of predictors increases, the training MSE consistently decreases until it is essentially 0. However, the testing MSE actually increases as the number of predictors in the model increase and even spikes extremely high at around 21 predictors. This outcome makes sense as the two df.train and df.test are completely unrelated (both are simply 25 variables of 25 random samples around a standard normal distribution) and thus the model trained heavily on df.train is unable to generalize effectively when used on the different df.test data frame.