tw2906_HW1

Tong Wu

2022-09-24

Problem 1

(a)

For the first given equation, it can be re-written as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i = n\overline{X}$$

Then for the second given equation:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n}\overline{X}X_i + \sum_{i=1}^{n}\overline{X}^2$$

$$\sum_{i=1}^{n} X_i^2 = (n-1)s^2 + 2\overline{X}\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\overline{X}^2$$

$$\sum_{i=1}^{n} X_i^2 = (n-1)s^2 + 2n\overline{X}^2 - n\overline{X}^2$$

$$\sum_{i=1}^{n} X_i^2 = (n-1)s^2 + n\overline{X}^2$$

```r
knitr::include_graphics("./src/1a.jpg")
```

P1.

(a) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i = n\bar{x}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$(n-1)S^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$(n-1)S^2 = \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n}\bar{x}x_i + \sum_{i=1}^{n}\bar{x}^2$$

$$\sum_{i=1}^{n} x_i^2 = (n-1)S^2 + 2\bar{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}^2$$

$$\sum_{i=1}^{n} x_i^2 = (n-1)S^2 + 2n\bar{x}^2 - n\bar{x}^2$$

$$\sum_{i=1}^{n} x_i^2 = (n-1)S^2 + n\bar{x}^2$$

Figure 1: Problem 1(a) solution

(b)

Since the answer from (a) shows that

$$\sum_{i=1}^{n} X_i^2 = (n-1)S^2 + n\overline{X}^2$$

then it can be calculated:

$$\begin{aligned}
\mathbb{E}nX_1^2 &= \mathbb{E}(n-1)S^2 + \mathbb{E}n\overline{X}^2 \\
\mathbb{E}(n-1)S^2 &= \mathbb{E}nX_1^2 - \mathbb{E}n\frac{(\sum_{i=1}^{n} X_i)^2}{n_2} \\
\mathbb{E}(n-1)S^2 &= \mathbb{E}nX_1^2 - \mathbb{E}X_1^2 + (n-1)X_1 \\
\mathbb{E}(n-1)S^2 &= \mathbb{E}(n-1)(X_1^2 + X_1) \\
\mathbb{E}S^2 &= \sigma^2
\end{aligned}$$

```r
knitr::include_graphics("./src/1b.jpg")
```

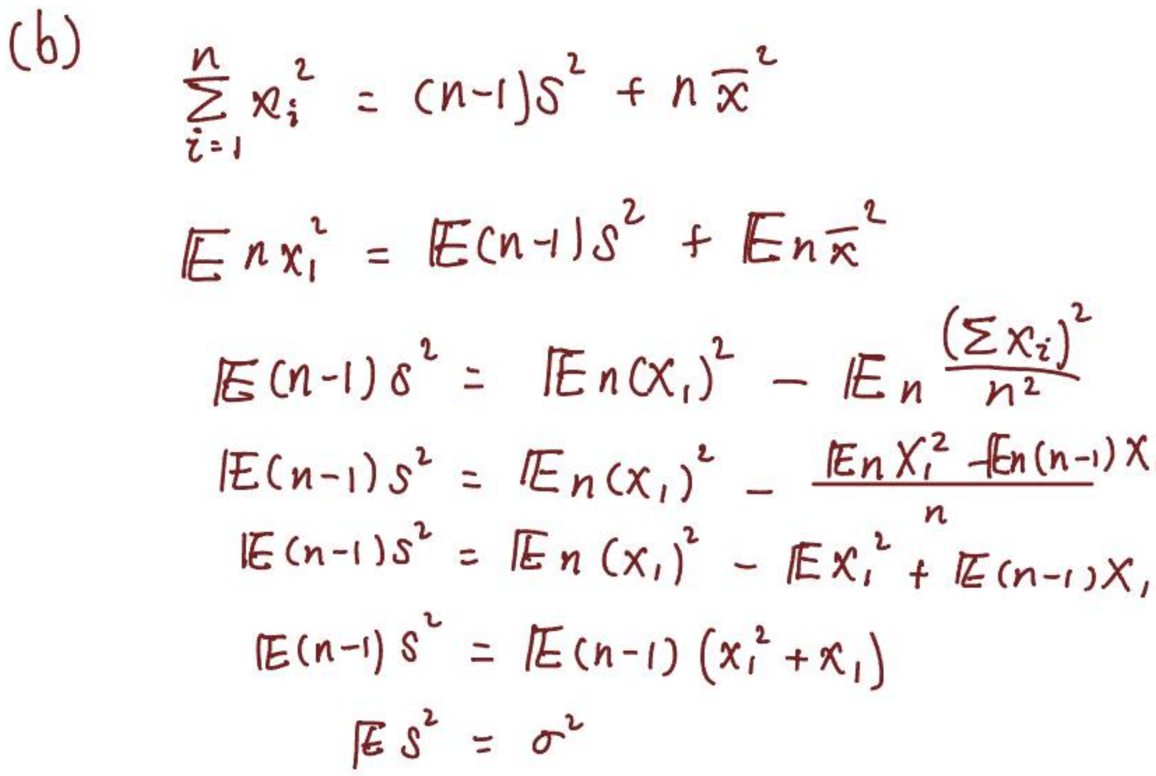


Figure 2: Problem 1(b) solution

(c)

Since the $\overline{X}$ and $X_i - \overline{X}$ are normal, and results from (b), can deduce that:

$$\overline{X} = 1/n \sum_{i=1}^{n} X_i \sim \mathbb{N}(\mu, \sigma^2/n)$$
$$X_i - \overline{X} \sim \mathbb{N}(0, ((n-1)/n)\sigma^2)$$
$$Cov(\overline{X}, X_i - \overline{X}) = Cov(\overline{X}, X_i) - Var(\overline{X})$$

```
knitr::include_graphics("./src/1c.jpg")
```

(C) Since the $\bar{X}$ and $X_i - \bar{X}$ are both normal, also, part (b) shows that $X_1, X_2, \ldots, X_n$ are independent and i.i.d.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$
$$X_i - \bar{X} \sim N(0, \frac{n-1}{n} \sigma^2)$$
$$Cov(\bar{X}, X_i - \bar{X}) = Cov(\bar{X}, X_i) - Var(\bar{X})$$
$$Cov(\bar{X}, X_i - \bar{X}) = \frac{1}{n} Cov(X_2 + \sum X_i X_j) - \frac{\sigma^2}{n}$$
$$Cov(\bar{X}, X_i - \bar{X}) = \frac{1}{n} Var(X_i - \frac{\sigma^2}{n})$$
$$Cov(\bar{X}, X_i - \bar{X}) = 0.$$

Which can be shown that, $\bar{X}$ is independent of $X_i - \bar{X}$. when $X_i, X_i - \bar{X}$ are both normal.

Figure 3: Problem 1(c) solution

(d)

Since the solution of 1(a) prove that

$$\sum_{i=1}^{n} X_i^2 = (n-1)s^2 + n\overline{X}^2$$

In 1(b) it proves that $X_i^2 - \overline{X}^2$ is independent of the $\overline{X}$, so the $\overline{X}$ should also independent to S^2

Problem 2

Solution for problem 2

```r
knitr::include_graphics("./src/2.jpg")
```

Problem 2

For the single predicoor, linear regression has model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{since } \bar{y} = \bar{x} = 0,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}, \quad \hat{\beta}_0 = \bar{y} - \frac{x_i \sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = 0.$$

$$\bar{y} = \hat{\beta}_1 \bar{x} \Rightarrow \sum (\hat{y}_i - \bar{y})^2 = \sum (x_i \hat{\beta}_1)^2$$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (x_i \hat{\beta}_1)^2}{\sum y_i^2} = \left[\frac{\left(\sum x_i y_i\right)^2}{\left(\sum x_i^2\right)^2} \sum_{i=1}^{n} x_i^2 \right] / \left(\sum y_i^2\right)$$

$$= \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2} / \sum y_i^2$$

$$= \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2} = r^2$$

Figure 4: Problem 2 solution

Problem 3

(a)

```r
set.seed(1)
x <- rnorm(100)
```

(b)

```r
eps <- rnorm(100, 0, sqrt(0.25))
```

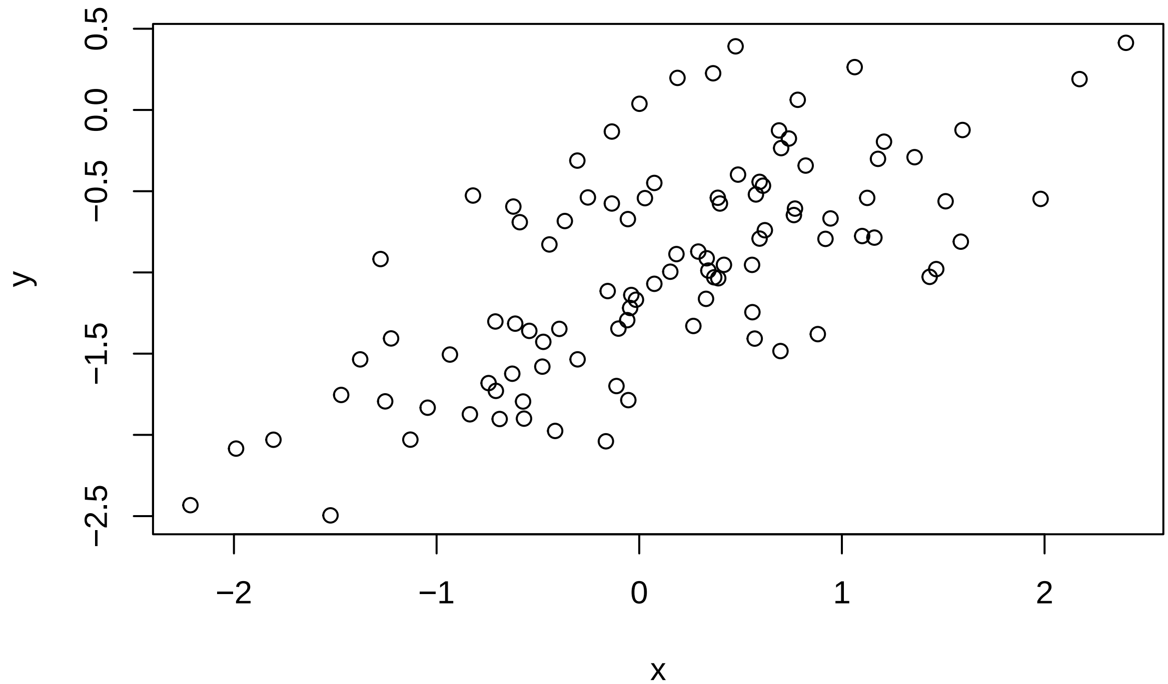
(c)

```r
y <- -1 + 0.5*x +eps
length(y)
```

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

(d)

```r
plot(xlab="x", ylab="y", x, y)
```



It shows the relationship regarding to y and x, which should be a linear function.

(e)

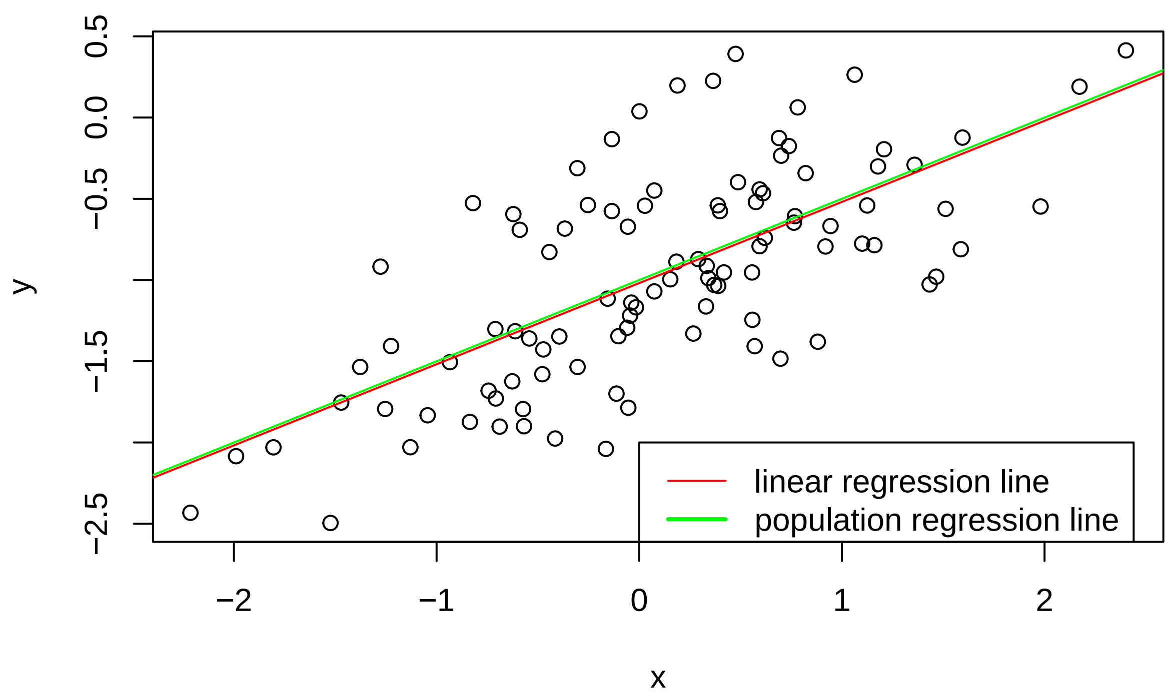
```
summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

The original coefficient β_0 and β_1 is -1 and 0.5The new $\hat{\beta}_0$ is -1.01885, $\hat{\beta}_1$ is 0.49947. The p-value is $4.853e^{-15}$, which is under the threshold value 0.05, so in this case the null hypothesis is more likely to be rejected.

(f)

```
plot(x, y)
abline(lm(y~x), col="red")
abline(-1, 0.5, col="green")
legend(0, -2, legend = c("linear regression line", "population regression line"),
       col=c("red","green"),lwd=1:2)
```



(g)

```
summary(lm(y~x+I(x^2)))
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.98252 -0.31270 -0.06441  0.29014  1.13500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
## x            0.50858    0.05399   9.420  2.4e-15 ***
## I(x^2)      -0.05946    0.04238  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 97 degrees of freedom
## Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
## F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

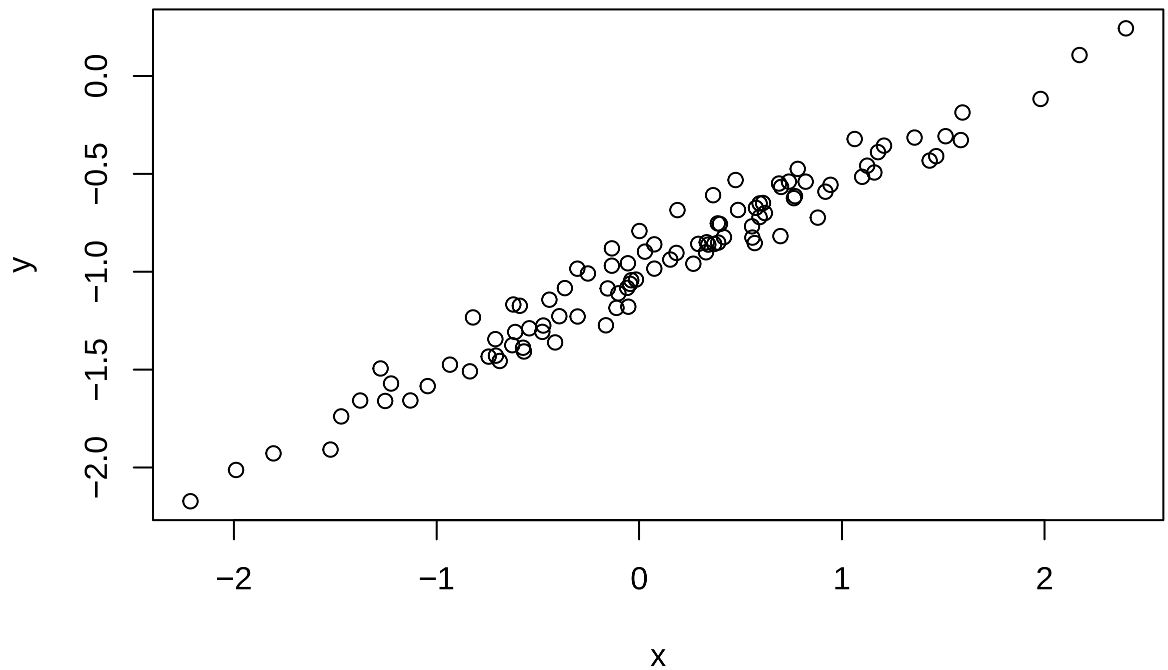
The R-squared has an increase than before, so it can be seen as the fit has been improved after adding the quadratic term.

(h)

```r
set.seed(1)
x_ln <- rnorm(100)
# change variable to 0.1, less noise
eps_ln <- rnorm(100, 0, 0.1)
y_ln <- -1 + 0.5*x_ln +eps_ln
length(y_ln)
```

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

```r
plot(xlab="x", ylab="y", x_ln, y_ln)
```

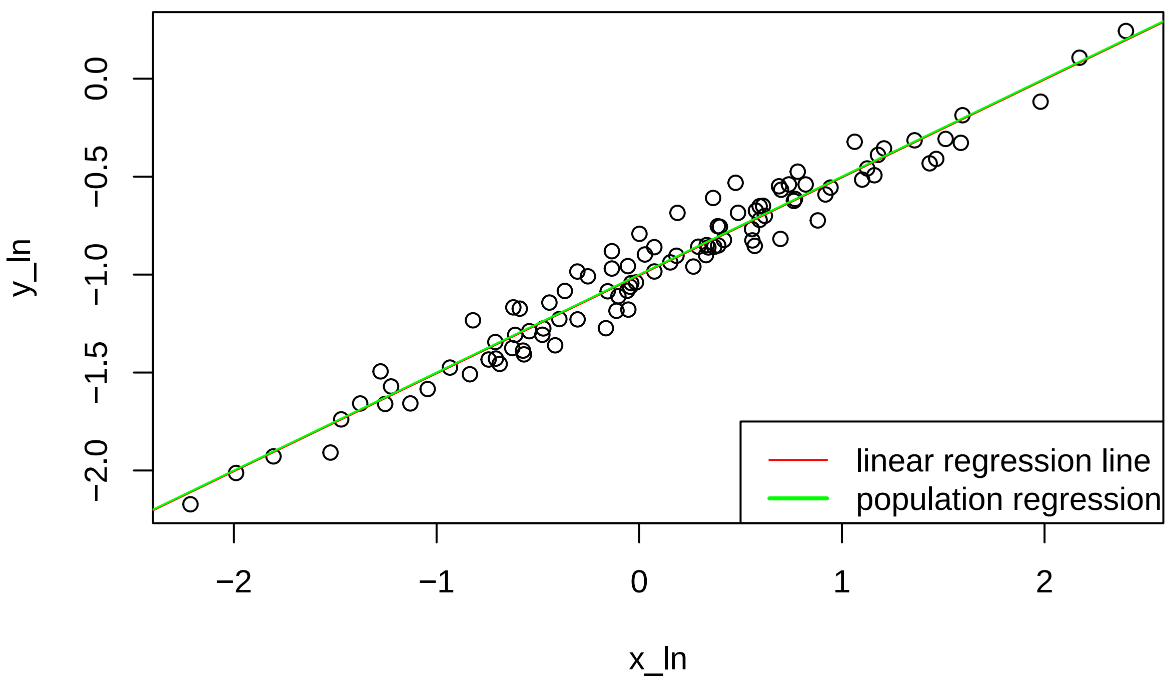



```
summary(lm(y_ln~x_ln))
```

```
##
## Call:
## lm(formula = y_ln ~ x_ln)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18768 -0.06138 -0.01395  0.05394  0.23462
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.003769   0.009699  -103.5   <2e-16 ***
## x_ln         0.499894   0.010773    46.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09628 on 98 degrees of freedom
## Multiple R-squared:  0.9565, Adjusted R-squared:  0.956
## F-statistic:  2153 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x_ln, y_ln)
abline(lm(y_ln~x_ln), col="red")
abline(-1, 0.5, col="green")
legend(0.5, -1.75
```

```
       , legend = c("linear regression line", "population regression line"),
       col=c("red","green"),lwd=1:2)
```



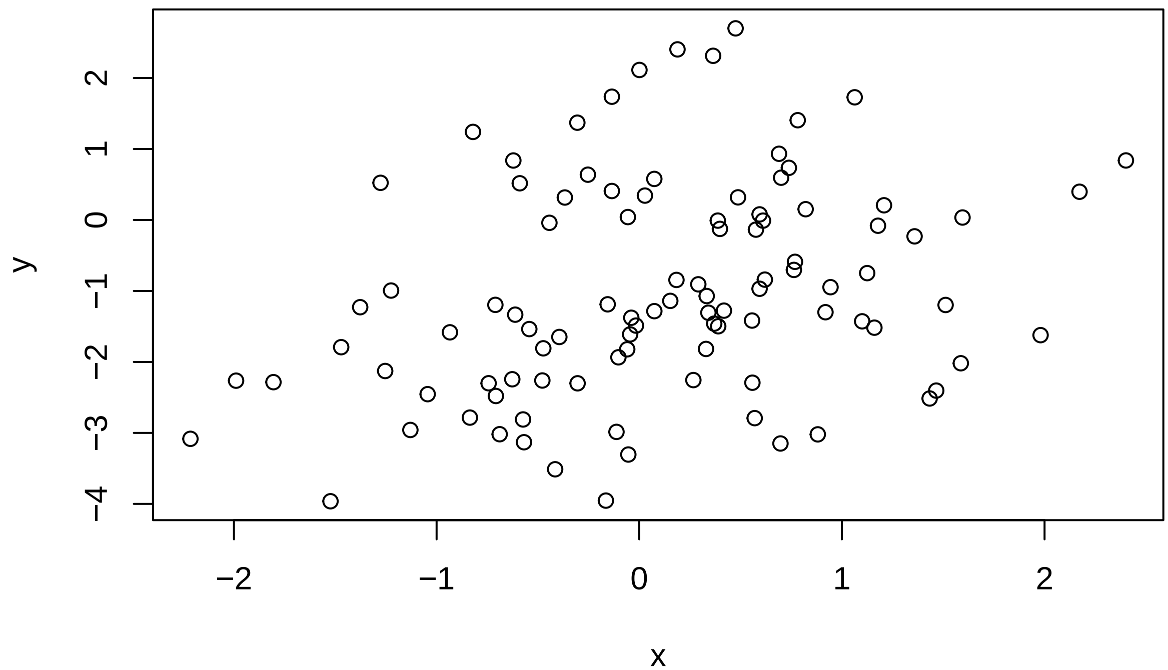
By changing the variance of the normal distribution to generate noise from 0.5 to 0.1, the Residual standard error has been decreased to 0.09628, and the R^2 increased to 0.9565. And the fitting graph shows that the regression line is more fit to the scattered points. These indicate that the model has less noise and the model is more fit.

(i)

```
set.seed(1)
x_mn <- rnorm(100)
# change variable to 1.5, make noisier
eps_mn <- rnorm(100, 0, 1.5)
y_mn <- -1 + 0.5*x_mn +eps_mn
length(y_mn)
```

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

```
plot(xlab="x", ylab="y", x_mn, y_mn)
```

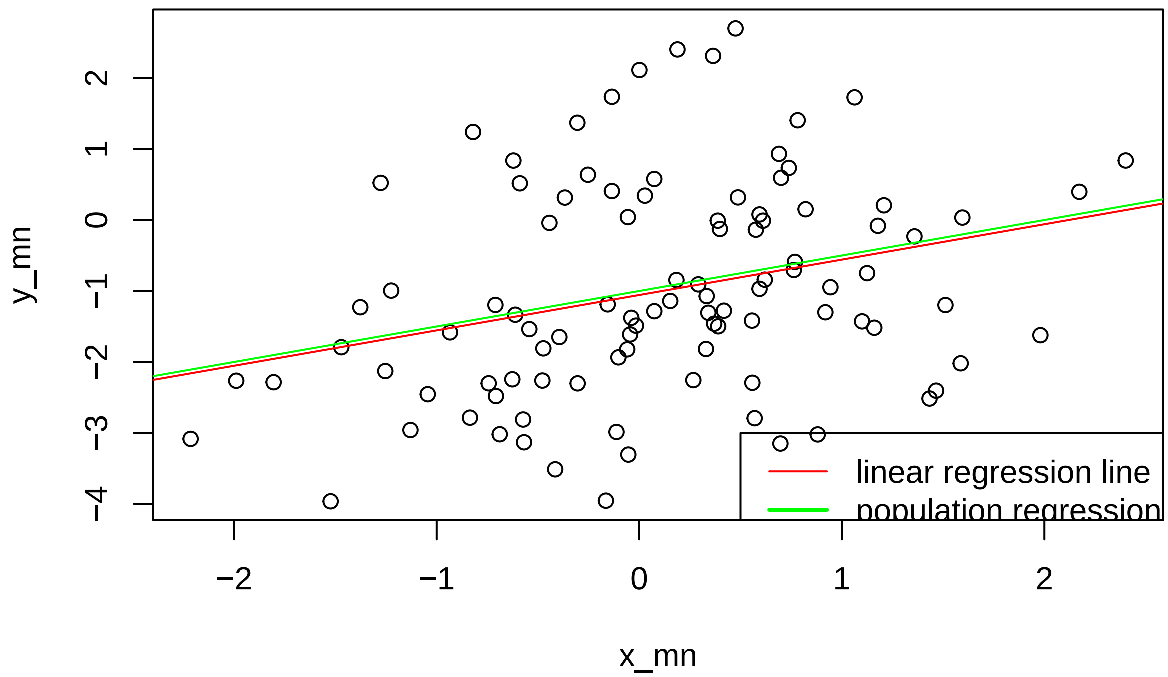



```
summary(lm(y_mn~x_mn))
```

```
##
## Call:
## lm(formula = y_mn ~ x_mn)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8153 -0.9206 -0.2092  0.8091  3.5193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0565     0.1455  -7.262 9.16e-11 ***
## x_mn          0.4984     0.1616   3.084  0.00265 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.444 on 98 degrees of freedom
## Multiple R-squared:  0.08849,    Adjusted R-squared:  0.07919
## F-statistic: 9.514 on 1 and 98 DF,  p-value: 0.00265
```

```
plot(x_mn, y_mn)
abline(lm(y_mn~x_mn), col="red")
abline(-1, 0.5, col="green")
```

```r
legend(0.5, -3, legend = c("linear regression line", "population regression line"),
       col=c("red","green"),lwd=1:2)
```



By changing the noise variance to 1.5, the Residual standard error increased to 1.444, and the R^2 decreased to 0.08849. For the diagram, the regression line is more flat and most part of scattered point are not fitted well. These indicate that this model has more noise which results the not well fit.

(j)

```r
# original data set
confint(lm(y~x), level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```r
# less noisy data set
confint(lm(y_ln~x_ln), level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0230161 -0.9845224
## x_ln         0.4785159  0.5212720
```

```r
# noisier data set
confint(lm(y_mn~x_mn), level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.3452411 -0.7678367
## x_mn         0.1777382  0.8190806
```

The less noise that the model has, the interval will be more narrow. For example, with the less noisy data set has more close to the original β_0 and β_1, and the noisier data set has bigger difference.

Problem 4

Sales onto newspaper

```r
Advertising <- read.csv("./src/Advertising.csv",sep=',')
str(Advertising)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
##  $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
##  $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
##  $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

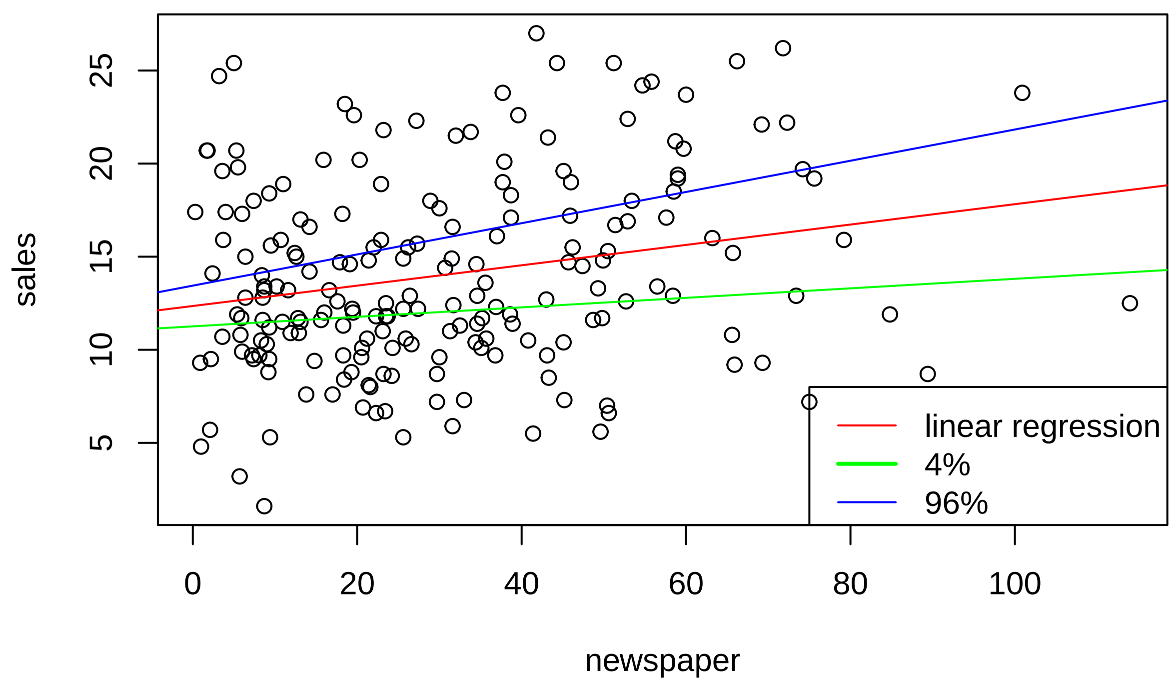
```r
#Sales onto Newspaper
lm.sales_newspaper <- lm(Advertising$sales~Advertising$newspaper)
summary(lm.sales_newspaper)
```

```
##
## Call:
## lm(formula = Advertising$sales ~ Advertising$newspaper)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           12.35141    0.62142   19.88  < 2e-16 ***
## Advertising$newspaper  0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

```
plot(Advertising$newspaper, Advertising$sales, xlab = "newspaper", ylab = "sales" )
abline(lm.sales_newspaper, col="red")
confint(lm.sales_newspaper, level=0.92)
```

```
##                                4 %        96 %
## (Intercept)             11.25788302 13.44493112
## Advertising$newspaper    0.02552451  0.08386169
```

```
abline(coef=confint(lm.sales_newspaper, level=0.92)[,1], col="green")
abline(coef=confint(lm.sales_newspaper, level=0.92)[,2], col="blue")
legend(75, 8, legend = c("linear regression line", "4%", "96%"),
       col=c("red","green", "blue"),lwd=1:2)
```



Sales onto TV

```
#Sales onto TV
lm.sales_TV <- lm(Advertising$sales~Advertising$TV)
summary(lm.sales_TV)
```

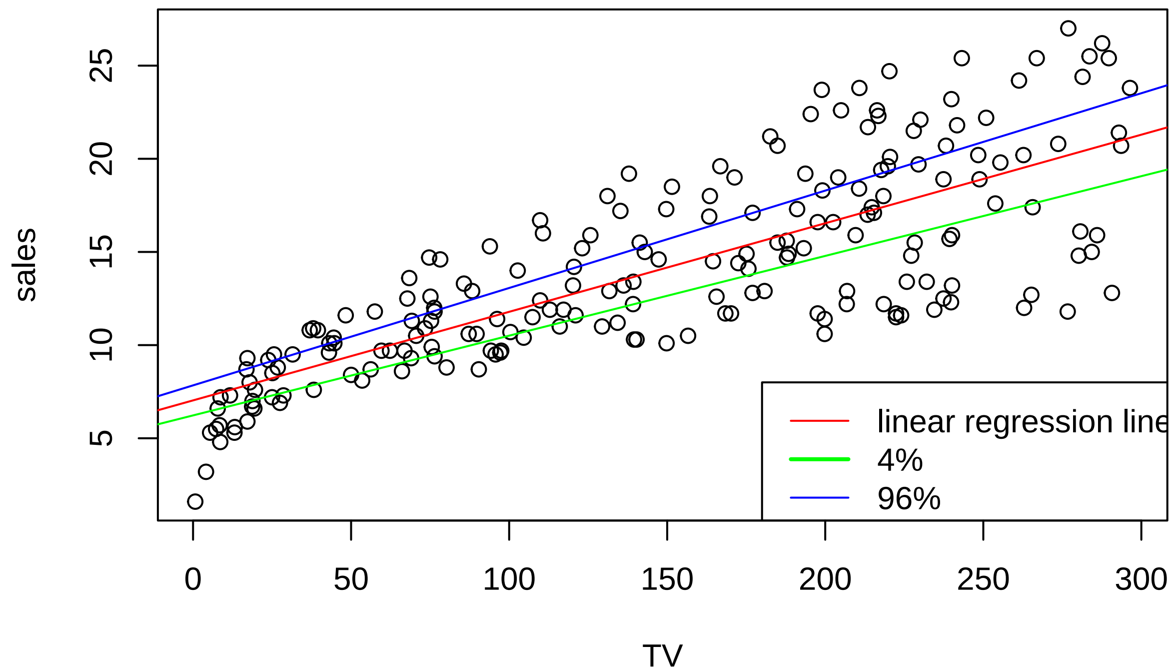
```
##
## Call:
## lm(formula = Advertising$sales ~ Advertising$TV)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.032594   0.457843   15.36   <2e-16 ***
## Advertising$TV 0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
plot(Advertising$TV, Advertising$sales, xlab = "TV", ylab = "sales" )
abline(lm.sales_TV, col="red")
confint(lm.sales_TV, level=0.92)
```

```
##                       4 %       96 %
## (Intercept)    6.22691926 7.83826784
## Advertising$TV 0.04280193 0.05227135
```

```
abline(coef=confint(lm.sales_TV, level=0.92)[,1], col="green")
abline(coef=confint(lm.sales_TV, level=0.92)[,2], col="blue")
legend(180, 8, legend = c("linear regression line", "4%", "96%"),
       col=c("red","green", "blue"),lwd=1:2)
```



Sales onto radio

```r
str(Advertising)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
##  $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
##  $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
##  $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

```r
#Sales onto radio
lm.sales_radio <- lm(Advertising$sales~Advertising$radio)
summary(lm.sales_radio)
```

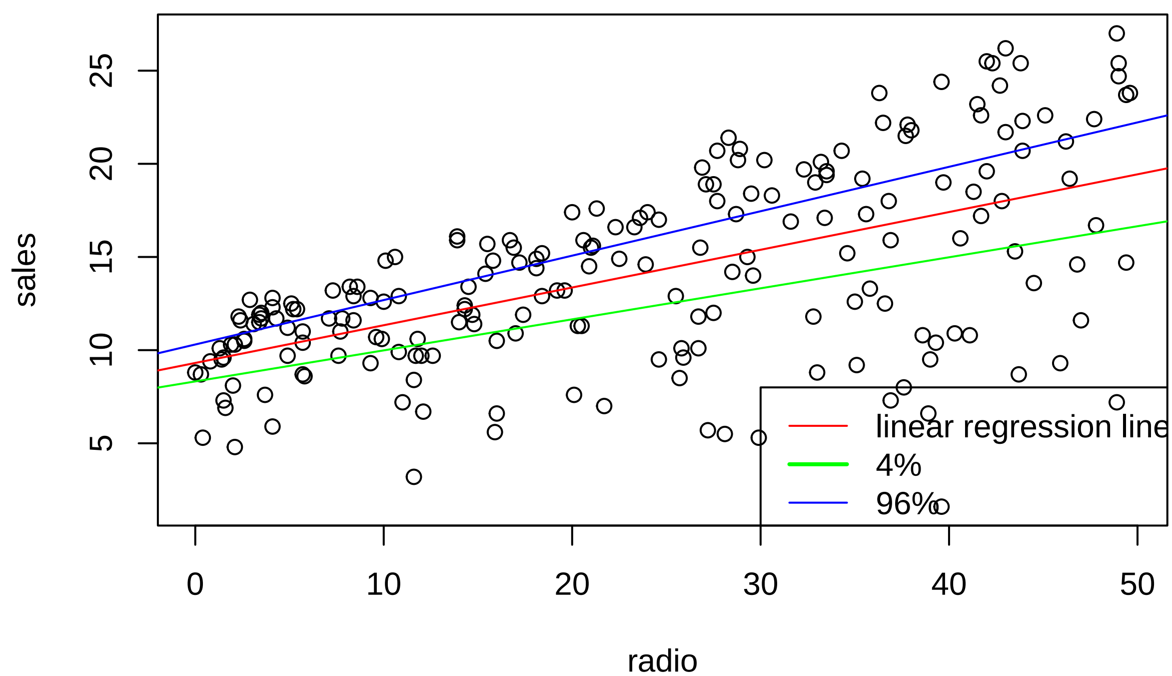
```
##
## Call:
## lm(formula = Advertising$sales ~ Advertising$radio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.31164    0.56290  16.542   <2e-16 ***
## Advertising$radio   0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
plot(Advertising$radio, Advertising$sales, xlab = "radio", ylab = "sales" )
abline(lm.sales_radio, col="red")
confint(lm.sales_radio, level=0.92)
```

```
##                         4 %       96 %
## (Intercept)       8.3210922 10.3021840
## Advertising$radio 0.1665776  0.2384139
```

```r
abline(coef=confint(lm.sales_radio, level=0.92)[,1], col="green")
abline(coef=confint(lm.sales_radio, level=0.92)[,2], col="blue")
legend(30, 8, legend = c("linear regression line", "4%", "96%"),
       col=c("red","green", "blue"),lwd=1:2)
```



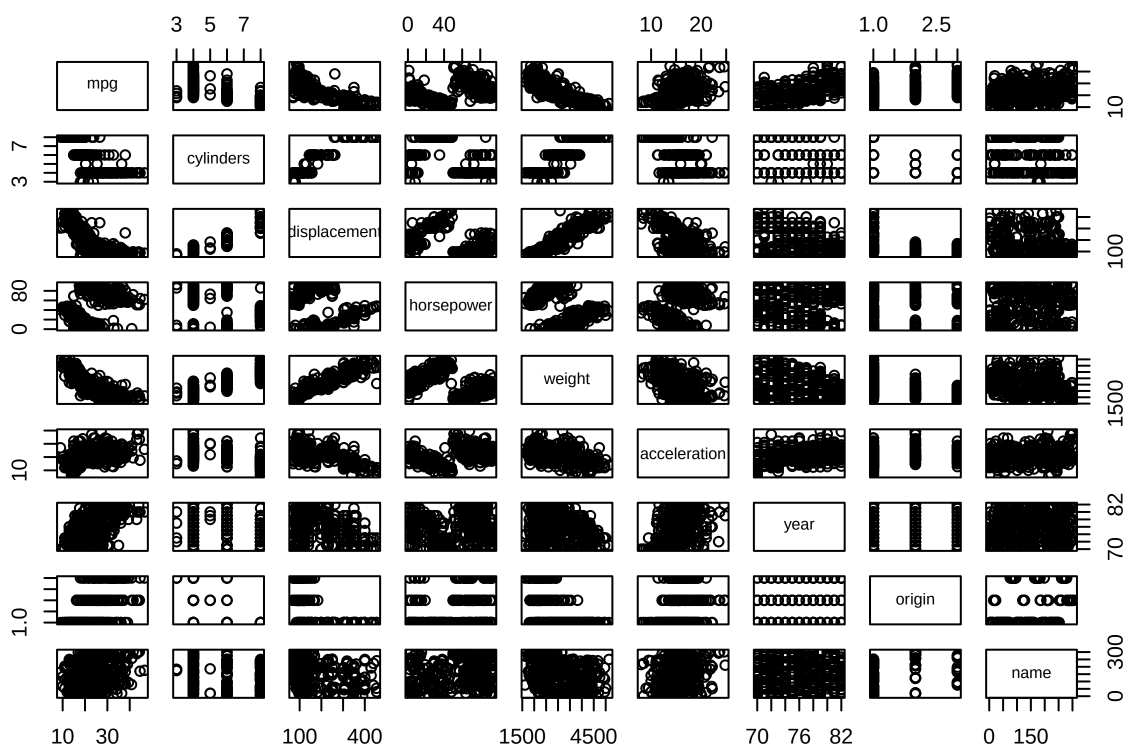
Problem 5

(a)

```
Auto <- read.csv("./src/Auto.csv",sep=',')
str(Auto)
```

```
## 'data.frame':    397 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : chr  "130" "165" "150" "150" ...
##  $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel
```

```
#head(Auto)
Auto[,4] = as.numeric(factor(Auto[,4]))
Auto[,9] = as.numeric(factor(Auto[,9]))
pairs(Auto[,1:9])
```



(b)

```r
cor(Auto[1:8])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7762599   -0.8044430  0.4228227 -0.8317389
## cylinders    -0.7762599  1.0000000    0.9509199 -0.5466585  0.8970169
## displacement -0.8044430  0.9509199    1.0000000 -0.4820705  0.9331044
## horsepower    0.4228227 -0.5466585   -0.4820705  1.0000000 -0.4821507
## weight       -0.8317389  0.8970169    0.9331044 -0.4821507  1.0000000
## acceleration  0.4222974 -0.5040606   -0.5441618  0.2662877 -0.4195023
## year          0.5814695 -0.3467172   -0.3698041  0.1274167 -0.3079004
## origin        0.5636979 -0.5649716   -0.6106643  0.2973734 -0.5812652
##              acceleration       year     origin
## mpg             0.4222974  0.5814695  0.5636979
## cylinders      -0.5040606 -0.3467172 -0.5649716
## displacement   -0.5441618 -0.3698041 -0.6106643
## horsepower      0.2662877  0.1274167  0.2973734
## weight         -0.4195023 -0.3079004 -0.5812652
## acceleration    1.0000000  0.2829009  0.2100836
## year            0.2829009  1.0000000  0.1843141
## origin          0.2100836  0.1843141  1.0000000
```

(c)

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

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Coefficients:
##  (Intercept)     cylinders  displacement    horsepower        weight
##   -21.284403     -0.292654      0.016034      0.007942     -0.006870
## acceleration          year        origin
##     0.153913      0.773442      1.346437
```

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

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.629 -2.034 -0.046  1.801 13.010
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -2.128e+01  4.259e+00  -4.998 8.78e-07 ***
## cylinders    -2.927e-01  3.382e-01  -0.865   0.3874    
## displacement  1.603e-02  7.284e-03   2.201   0.0283 *  
## horsepower    7.942e-03  6.809e-03   1.166   0.2442    
## weight       -6.870e-03  5.799e-04 -11.846  < 2e-16 ***
## acceleration  1.539e-01  7.750e-02   1.986   0.0477 *  
## year          7.734e-01  4.939e-02  15.661  < 2e-16 ***
## origin        1.346e+00  2.691e-01   5.004 8.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.331 on 389 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8188 
## F-statistic: 256.7 on 7 and 389 DF,  p-value: < 2.2e-16
```

i

Since the p-value is smaller than the threshold value 0.05, so the null hypothesis should be rejected. So it can be said that there is a relationship between the predictors and the response.

ii

The displacement, weight, acceleration, year and origin have a statistically significant relationship to the response.

iii

It suggests that in each year, the mpg will increase 0.77, which implies that the horse power will increase and the cylinders will decrease in each year.

(d)

```r
lm.fit_5d = lm(sqrt(Auto$mpg)~log(Auto$cylinders)+sqrt(Auto$displacement)+
                 (Auto$horsepower^2)+log(Auto$weight)+sqrt(Auto$acceleration)+
                 (Auto$year)+(Auto$origin^2))
summary(lm.fit_5d)
```

```
## 
## Call:
## lm(formula = sqrt(Auto$mpg) ~ log(Auto$cylinders) + sqrt(Auto$displacement) + 
##     (Auto$horsepower^2) + log(Auto$weight) + sqrt(Auto$acceleration) + 
##     (Auto$year) + (Auto$origin^2))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.99376 -0.16798  0.00305  0.16405  1.01267 
## 
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             14.8106443  1.0946491  13.530  < 2e-16 ***
```

```
## log(Auto$cylinders)      -0.0822049  0.1708188  -0.481  0.63062
## sqrt(Auto$displacement)  0.0140099  0.0206399   0.679  0.49768
## Auto$horsepower          0.0007579  0.0005997   1.264  0.20708
## log(Auto$weight)        -2.0736757  0.1570709 -13.202  < 2e-16 ***
## sqrt(Auto$acceleration)  0.0759327  0.0539977   1.406  0.16046
## Auto$year                0.0786438  0.0043057  18.265  < 2e-16 ***
## Auto$origin              0.0667427  0.0246443   2.708  0.00706 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2911 on 389 degrees of freedom
## Multiple R-squared:  0.8713, Adjusted R-squared:  0.869
## F-statistic: 376.1 on 7 and 389 DF,  p-value: < 2.2e-16
```

From the data, null hypothesis should be rejected, since the p-value is smaller than the threshold value 0.05. The residual standard error decreased and R^2 has a slightly increase.

Problem 6

$$\overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 0.4276$$

$$\overline{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 19.91$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{20}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{20}(x_i - \overline{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{20}(y_i x_i - \overline{y}x_i - \overline{x}y_i + \overline{xy})}{\sum_{i=1}^{20}(x_i^2 - 2x_i\overline{x} + \overline{x}^2)}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{20} y_i x_i - \overline{y}\sum_{i=1}^{20} x_i - \overline{x}\sum_{i=1}^{20} y_i + \overline{xy}}{\sum_{i=1}^{20} x_i^2 - \overline{x}\sum_{i=1}^{20} 2x_i + \overline{x}^2}$$

$$\hat{\beta}_1 = \frac{216.6 - 398.2 * 0.4276 - 19.91 * 8.552 + 19.91 * 0.4276 * 20}{5.196 + 0.4276^2 * 20 - 2 * 8.552 * 0.4276}$$

$$\hat{\beta}_1 = 30.1$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 19.91 - 30.1 * 0.4276 = 7.039$$

So,

$$\hat{y} = 30.1x + 7.039$$

, and when $x = 0.5$,

$$\hat{y} = 22.089$$

$$R^2 = \frac{\sum_{i=1}^{20}(30.1x_i + 7.039 - \overline{y})^2}{\sum_{i=1}^{20}(y_i^2 + \overline{y}^2 - 2y_i\overline{y})} = 0.98$$

$$\sigma^2 = \frac{1 - R^2 * TSS}{n - p - 1} = 1.85$$

```
knitr::include_graphics("./src/6.jpg")
```

Problem 6

$$\sum_{i=1}^{20} x_i = 8.552 \quad \Rightarrow \quad \bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 0.4276$$

$$\sum_{i=1}^{20} y_i = 398.2 \quad \Rightarrow \quad \bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 19.91$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{20} (y_i x_i - \bar{y} x_i - \bar{x} y_i + \bar{x}\bar{y})}{\sum_{i=1}^{20} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}$$

$$= \frac{216.6 - 19.91 \times 8.552 - 0.4276 \cdot 398.2 + 0.4276 \times 19.91 \times 20}{5.196 - 2 \times 0.4276 \times 8.552 + 0.4276^2 \times 20}$$

$$= 30.1$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 19.91 - 30.1 \cdot 0.4276$$
$$= 7.039$$

$$\hat{y} = 30.1x + 7.039, \text{ when } x = 0.5, \ \hat{y} = 22.089$$

$$R^2 = \frac{\sum_{i=1}^{20} (30.1 x_i + 7.039 - \bar{y})^2}{\sum_{i=1}^{20} (y_i^2 + \bar{y}^2 - 2y_i\bar{y})} = 0.98$$

$$\sigma^2 = \frac{1 - R^2 \cdot TSS}{n-p-1} = 1.85$$

Figure 5: Problem 6 solution

Problem 7

If the model accept the null hypothesis, then the equation is

$$\frac{TSS - RSS}{p} / \frac{RSS}{n-p-1} = \frac{11.62 - 8.95}{6} / \frac{8.95}{45-6-1} = 1.889$$

```r
pf(1.889, 6, 38, lower.tail=FALSE)
```

```
## [1] 0.1080044
```

p-value is 0.1080044