Patrick Rogan
UNI: psr2125

STAT4702 Homework 8

3.7.3
X_1 = GPA
X_2 = IQ
X_3 = Gender (1 for Female and 0 for Male)
X_4 = Interaction between GPA and IQ
X_5 = Interaction between GPA and Gender
$\hat{\beta}_0 = 50$
$\hat{\beta}_1 = 20$
$\hat{\beta}_2 = 0.07$
$\hat{\beta}_3 = 35$
$\hat{\beta}_4 = 0.01$
$\hat{\beta}_5 = -10$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_1 * x_2 + \hat{\beta}_5 x_1 * x_3$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 + \hat{\beta}_5 x_5 + (\hat{\beta}_4 x_1 + \hat{\beta}_2) x_2$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_5 x_5 + (\hat{\beta}_4 x_2 + \hat{\beta}_1) x_1$$

a. iii is correct, fixing GPA and IQ, we are interested in $\hat{\beta}_3 * x_3 + \hat{\beta}_5 * x_1 * x_2$ which, for a female is $35 + -10 * GPA$, for a GPA value of > 3.5 a male will earn more.

b. $\hat{y} = 50 + 20 * 4.0 + 0.07 * 110 + 35 * 1 + 0.01 * 440 + -10 * 4$
$= 137.1$

c. False. While this is suggestive that there isn't an interaction effect, more evidence should be gathered before dismissing the possibility of interaction. For example, the p-vale, confidence interval of β_4, and potentially hidden higher order interactions should all be examined before dismissing an interaction effect.

3.7.4

a. Given the true relationship is linear, one would expect the cubic model to produce a lower training RSS. The greater flexibility in the model, in general, should allow for a closer fit given expected random fluctuations in the data.

b. For the test RSS one would expect the linear model to have superior performance. The cubic model presumably would have fit features in the training data that one would not expect to be replicated in the test data.

c. This is similar to part a. One would expect a more flexible model to provide a better fit, but it is highly dependant on the sample.

d. For the test RSS it is impossible to say. For a highly nonlinear case the test RSS would be minimized for a cubic model. However, as the underlying behaviour of the model converges to a linear form, there is no way of knowing which model would provide a lower RSS.

3.7.5 See supporting work section.

3.7.8

a.

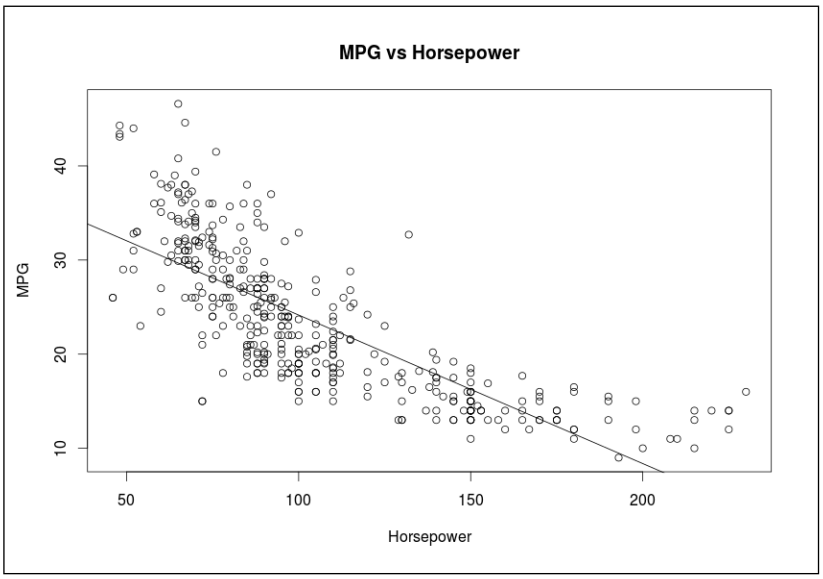
i. Yes, there is a relationship between the predictor and the response.

ii. The relationship appears somewhat strong as there is a 21% error associated with the RSE and an R^2 value of 0.606 indicating roughly 60% of MPG can be explained by horsepower.

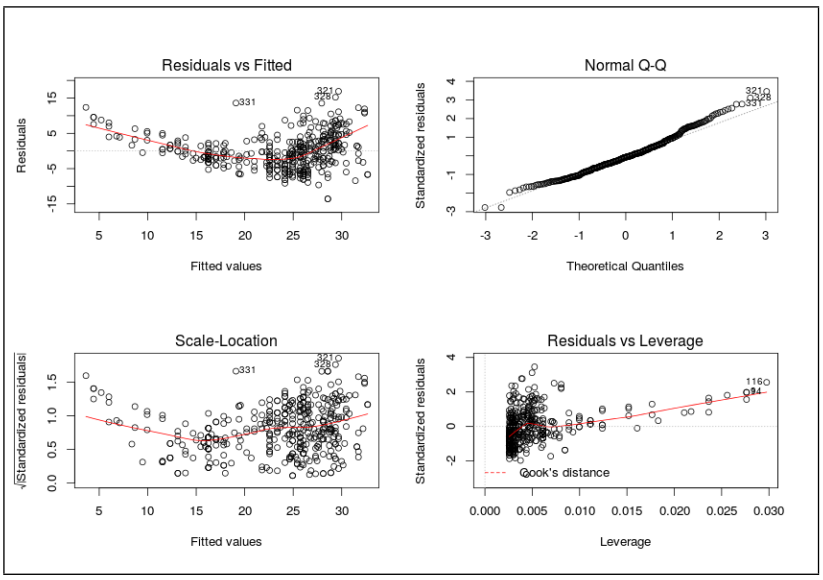
iii. The relationship between the predictor and the response is negative.

iv.
```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

b.

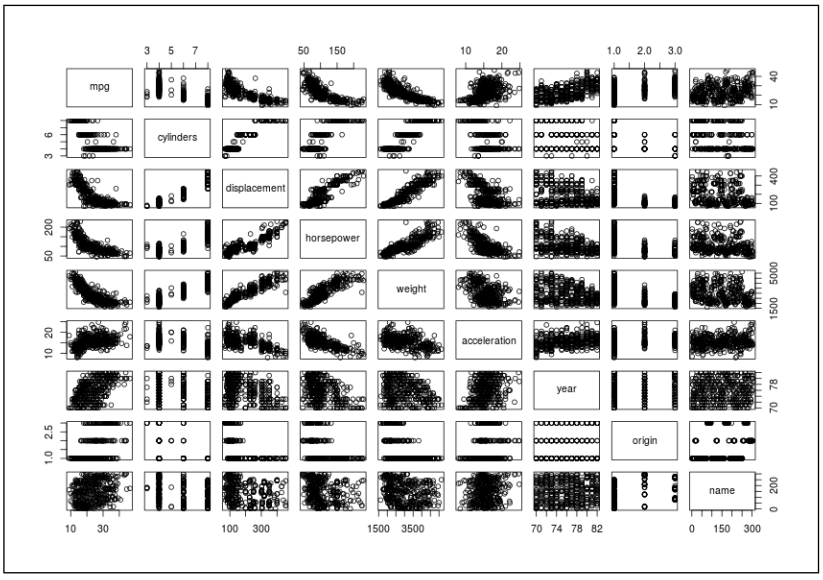


c. There is an obvious trend in the residual plot suggesting that a different model would produce a better fit. Additionally, there appears to be at least one outlier (331) and several high leverage points including 94 and 116.



3.7.9

a.



b.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1.0000	-0.7776	-0.8051	-0.7784	-0.8322	0.4233	0.5805	0.5652
cylinders	-0.7776	1.0000	0.9508	0.8430	0.8975	-0.5047	-0.3456	-0.5689
displacement	-0.8051	0.9508	1.0000	0.8973	0.9330	-0.5438	-0.3699	-0.6145
horsepower	-0.7784	0.8430	0.8973	1.0000	0.8645	-0.6892	-0.4164	-0.4552
weight	-0.8322	0.8975	0.9330	0.8645	1.0000	-0.4168	-0.3091	-0.5850
acceleration	0.4233	-0.5047	-0.5438	-0.6892	-0.4168	1.0000	0.2903	0.2127
year	0.5805	-0.3456	-0.3699	-0.4164	-0.3091	0.2903	1.0000	0.1815
origin	0.5652	-0.5689	-0.6145	-0.4552	-0.5850	0.2127	0.1815	1.0000

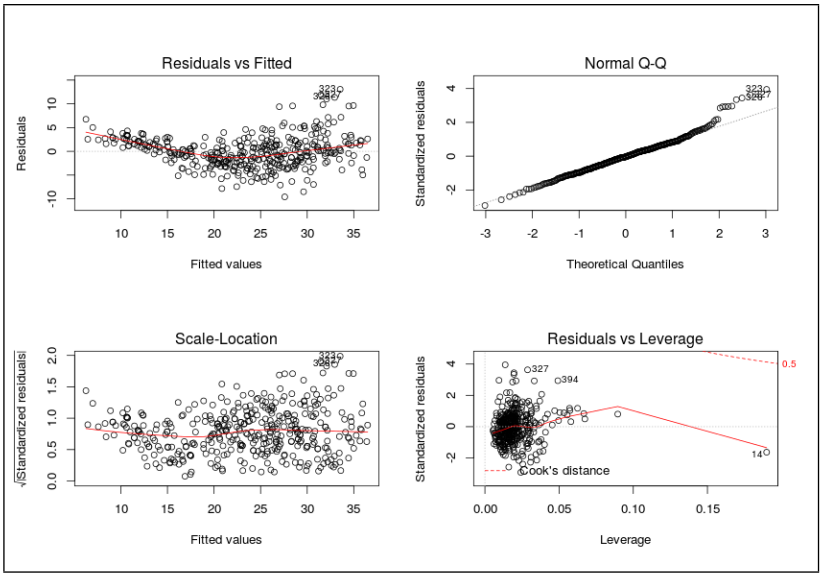
c.

i. Yes, the F-statistic is 252 (significantly greater than 1) with significant p-value and the R^2 value indicates that the linear model accounts for roughly 82% of the variation.

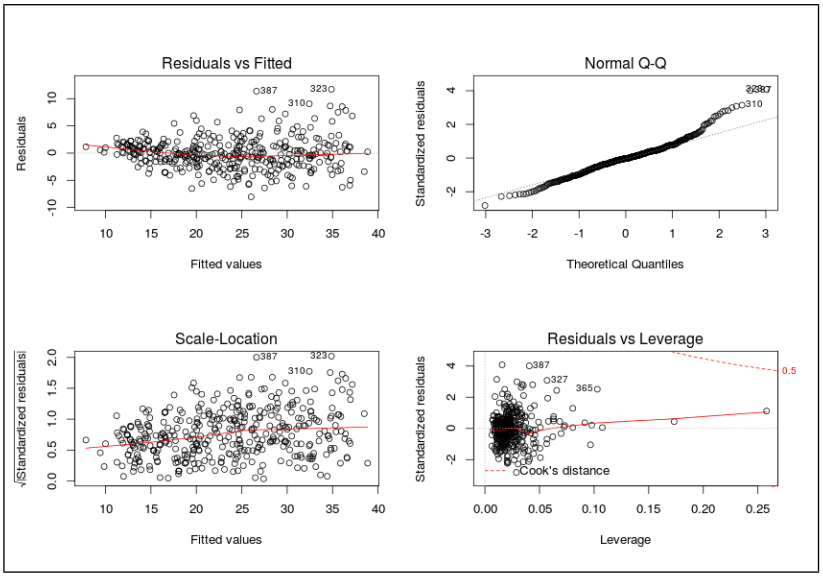
ii. Displacement, weight, year and origin all appear to have a statistically significantly relationship to the response.

iii. The statistically significant coefficient for year is fairly large with respect to the other coefficients and, removing year from the model and re-calculating R^2 we have $R^2_{w/age} = 0.821$ and $R^2_{wo/age} = 0.721$ indicating a significant portion of the model variation is accounted for by this predictor.

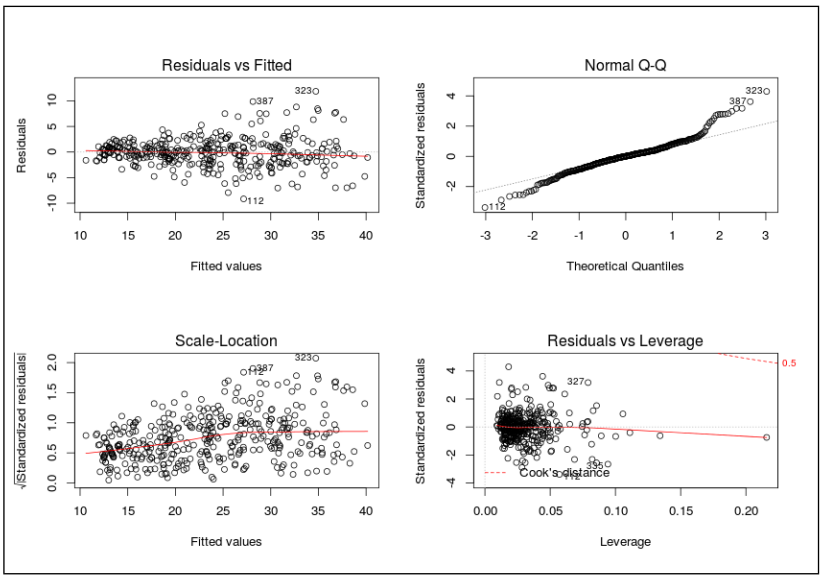
d. Looking at the residuals, there appears to be a pattern suggesting the model may need to be changed to relax the constraints to include interaction effects, higher order terms, or correlated terms. The leverage plot identifies point 14 as having particularly high leverage.



e. If we make the model as flexible as possible in terms of interaction effects (all combinations of interaction effects) then none of the terms appear significant. If, instead, we use a model that include interactions between horsepower and weight, and displacement and acceleration (pairs of variables selected from the correlation matrix) there are significant interactions between these variables. Additionally, if these terms are included, the R^2 value of the model increases from 0.821 in part d. to 0.865 and the trend in the Residuals vs Fitted diagnostic plot is less apparent.



f. Adding log terms for displacement, horsepower, and weight and a square term for year increases R^2 to 0.876 and nearly eliminates the trend in the Residuals vs Fitted diagnostic plot. That said the model is fairly complex and which may indicate overfitting of the data.



3.7.10

a. See supporting code.

b.

i. $\hat{\beta_0} = 13.04$, there is a statistically significant baseline when modelling sales based on Price, Urban and US.

ii. $\hat{\beta_1} = -0.054$, the coefficient for Price suggests a negative relationship between sales and price.

iii. $\hat{\beta_2} = -0.022$, given that factor Urban = 0 equates to "No", the store's presence in an urban location appears to have a negative impact on sales.

iv. $\hat{\beta_3} = 1.200$, given that factor US = 0 equates to "No", the store's presence in the US appears to have a positive impact on sales.

c. $\hat{y} = \hat{\beta_0} + \hat{\beta_1}x_1 + \hat{\beta_2}x_2 + \hat{\beta_3}x_3$

$$sales \approx \hat{\beta}_0 + \hat{\beta}_1 * Price + \begin{cases} 0 & if\ Not\ Urban\ and\ Not\ US \\ \hat{\beta}_2 & if\ Urban\ and\ not\ US \\ \hat{\beta}_3 & if\ Not\ Urban\ and\ US \\ \hat{\beta}_2 + \hat{\beta}_3 & if\ Urban\ and\ US \end{cases}$$

where
$\hat{\beta_0} = 13.04$
$\hat{\beta_1} = -0.054$
$\hat{\beta_2} = -0.022$
$\hat{\beta_3} = 1.200$

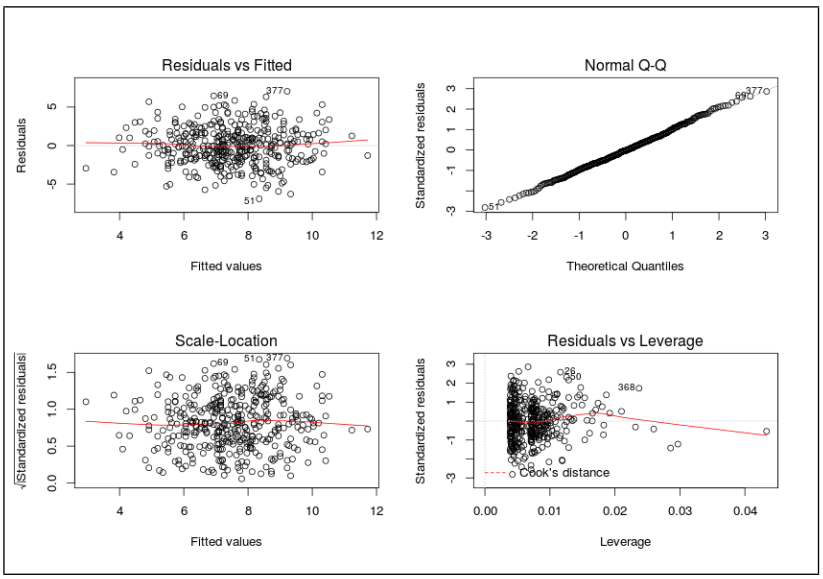
d. We can reject the null for Price and US.

e. See supporting code.

f. Both models fit the data about equally well. RSE values for both models are ≈ 2.47, $R^2 \approx 0.24$ and the F-test statistic is significant in both cases. That is, removing Urban from the model doesn't have much of an impact, and the model only accounts for about a quarter of the variability in sales.

g.
```
                2.5 %  97.5 %
(Intercept) 11.79032 14.2713
Price       -0.06476 -0.0442
USYes        0.69152  1.7078
```

h. Given the plot of Residuals vs Fitted, there does not seem to be any obvious outliers in the data. Given the average leverage $= \frac{p+1}{n} = \frac{3}{400} = 0.0075$ points to the right of point 368 in the Residuals vs Leverage plot should be investigated to determine an adverse effect on the model as they are high leverage points.



3.7.11

a. $\hat{\beta} = 1.994$, $SE = 0.106$, $t - statistic = 18.7$, $p - value =< 2e - 16$. This model amounts to a line of slope 2 with 0 intercept. Given the high t-statistic and resulting incredibly small p-value, the one predictor available is highly significant.

b. $\hat{\beta} = 0.3911$, $SE = 0.0209$, $t - statistic = 18.7$, $p - value =< 2e - 16$. Here the model is a line of slope 0.4 with 0 intercept. The t-statistic and p-value indicate the variable is statistically significant.

c. The coefficient and SE have changed, which are expected, but the t-statistic and p-value are the same.

d. See supporting work section for derivation, see supporting code section for R verification which produces a t-statistic value of 18.7, the same as what we had from before.

e. Clearly the result obtained in part d is the same if we exchange x and y, therefore the t-statistic for the regression of y onto x is the same as for x onto y in the 0 intercept case.

f. The results are the same, see supporting code for details.

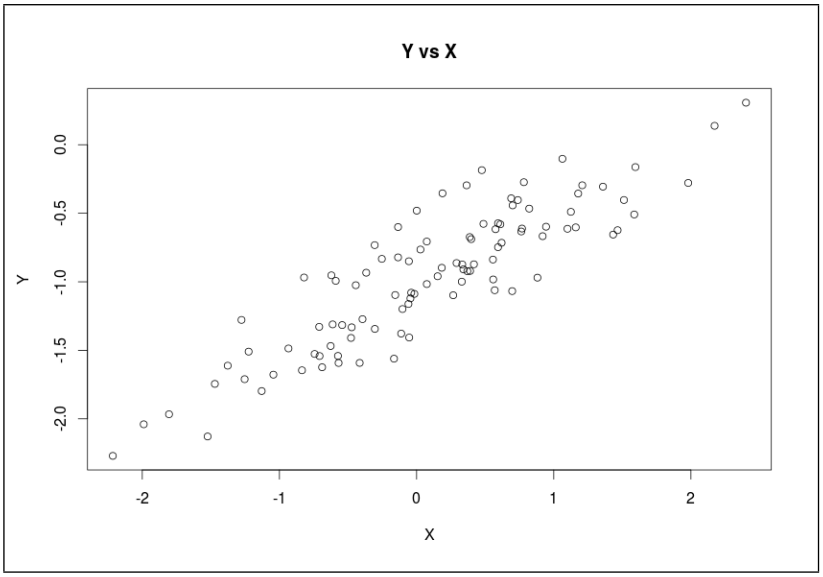
3.7.13

a. See supporting code.

b. See supporting code.

c. Length $= 100$, $\beta_0 = -1$, $\beta_1 = 0.5$

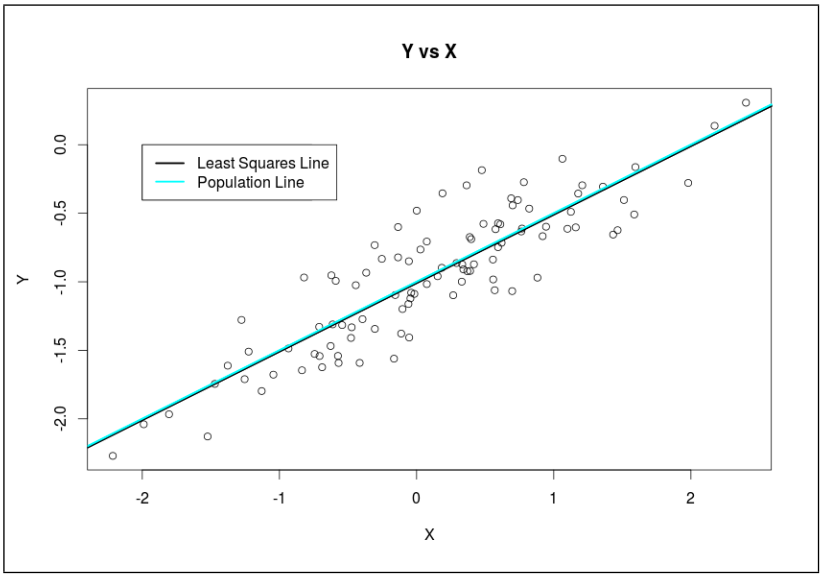
d. There is a linear relationship between X and Y with some noise. This is exactly what was required.



e. Note set.seed(1) is only set prior to setting X. If this value is set before any of the other random number generations the resulting Y will have perfectly linear data points.

$\hat{\beta}_0 = -1.0094$, $\hat{\beta}_1 = 0.4997$, that is, the estimated parameters are extremely close to the true parameter values.

f. Note the lines are basically right on top of each other. Given the printer produced black and white they may be difficult to distinguish. See supporting code section to confirm their existence.

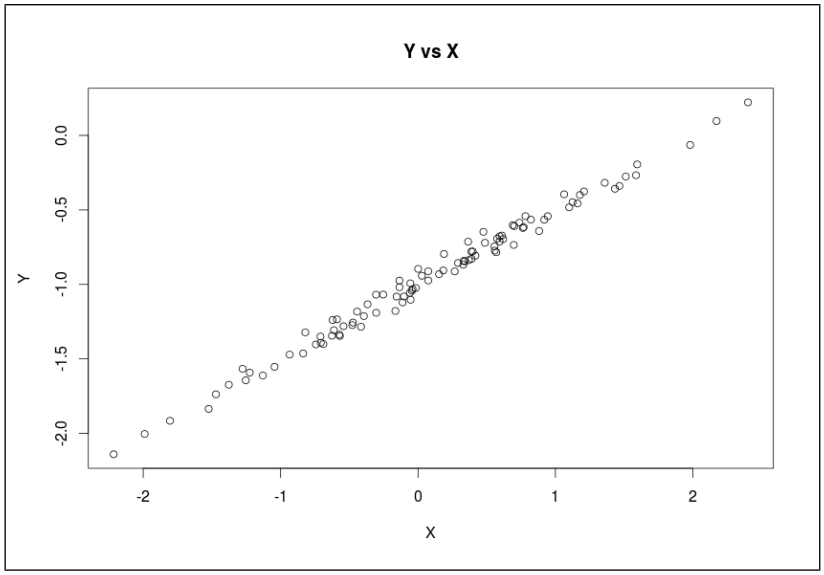


g. Adding a squared term increases R^2 by 0.005. The p-value for this term indicates that it is not statistically significant. So, while we have slightly improved out fit as is expected by a more flexible model, there is not a significant change.

h. For this part, the variance was set to 0.0025 (SD = 0.05).

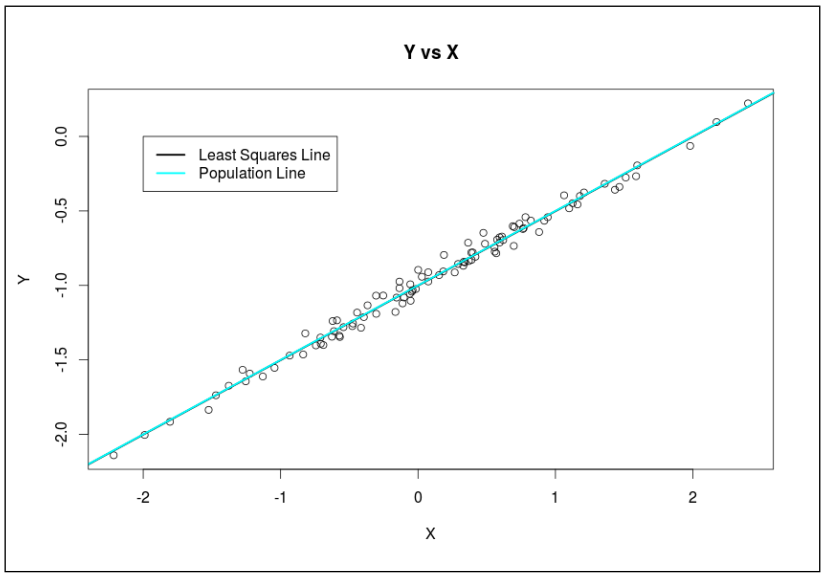
   c. Unchanged.

   d. The linear trend is much more pronounced than in the higher variance noise term case.



e. $\hat{\beta}_0 = -1.00188$, $\hat{\beta}_1 = 0.49995$ these values are $> 0.2\%$ different than the true values.

f.

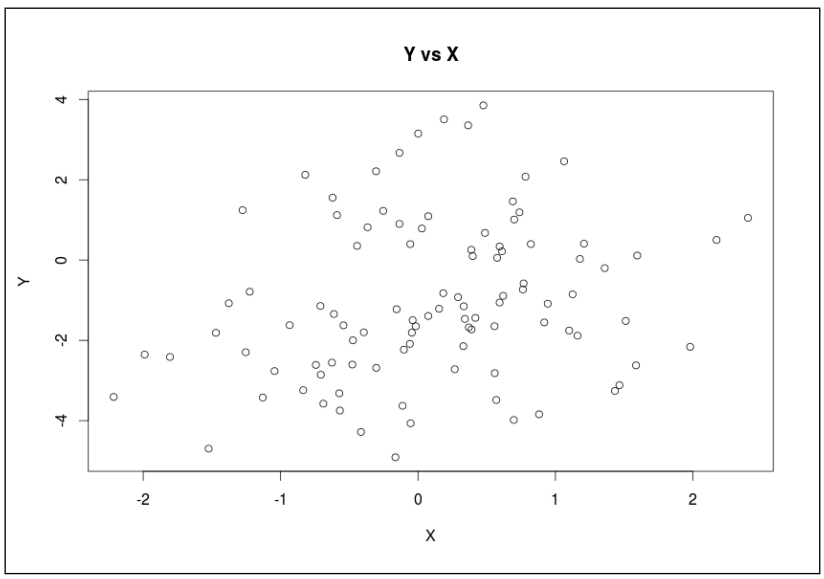


Clearly by reducing the variance in the error term the data produced are much closer to the true population line. Additionally, when fitting this data, the R^2 value is 0.989, indicating almost all of the variation is accounted for by the simple linear regression.

i. For this part, the variance was set to 4.0 (SD = 2.0).

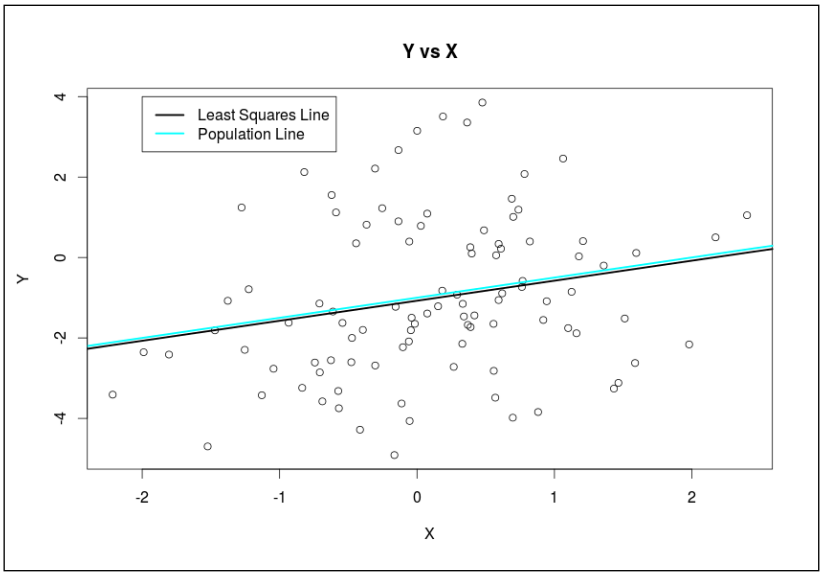
c. Unchanged.

d. The linear trend has been almost completely obfuscated by the variance introduced to the data.



e. $\hat{\beta}_0 = -1.075$, $\hat{\beta}_1 = 0.498$, even though there is significantly more variance in the data, the model coefficients are still fairly close to the true parameters.

f.



With the introduction of considerably more noise, the linear model still fits the data fairly well with the p-values for both coefficients 2.5e-07 and 0.023 for β_0 and β_1 respectively, still indicate significance. One significant change is the R^2 value has dropped substantially to 0.0517.

j. Original Model:

```
               2.5 %  97.5 %
(Intercept) -1.0575 -0.9613
X            0.4463  0.5532
```

Noisier Model:

```
               2.5 %  97.5 %
(Intercept) -1.46032 -0.6904
X            0.07032  0.9254
```

Less Noisy Model:

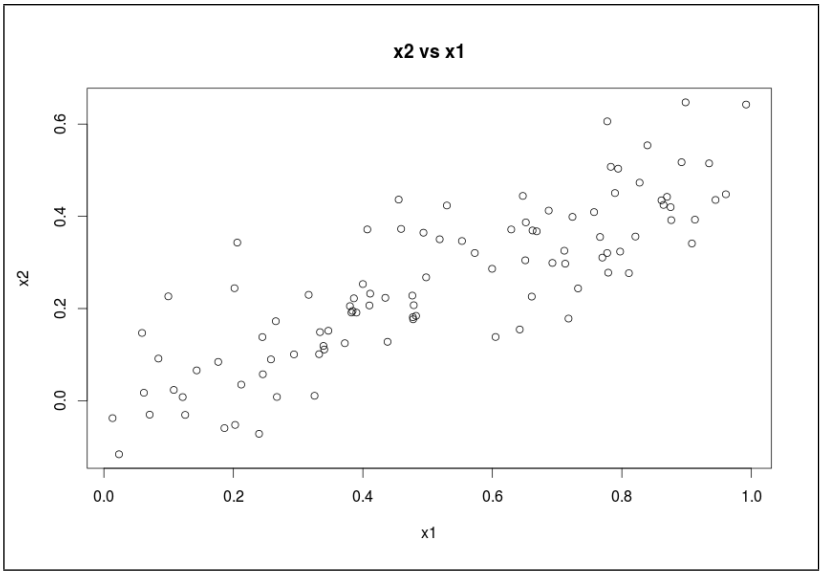
```
               2.5 %  97.5 %
(Intercept) -1.0115 -0.9923
X            0.4893  0.5106
```

The more noise in the model, the larger the 95% confidence intervals and the less sure we are what values the true parameters take.

3.7.14

a. $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon$ where
$\beta_0 = 2$
$\beta_1 = 2$
$\beta_2 = 0.3$

b. x1 and x2 are positively correlated.



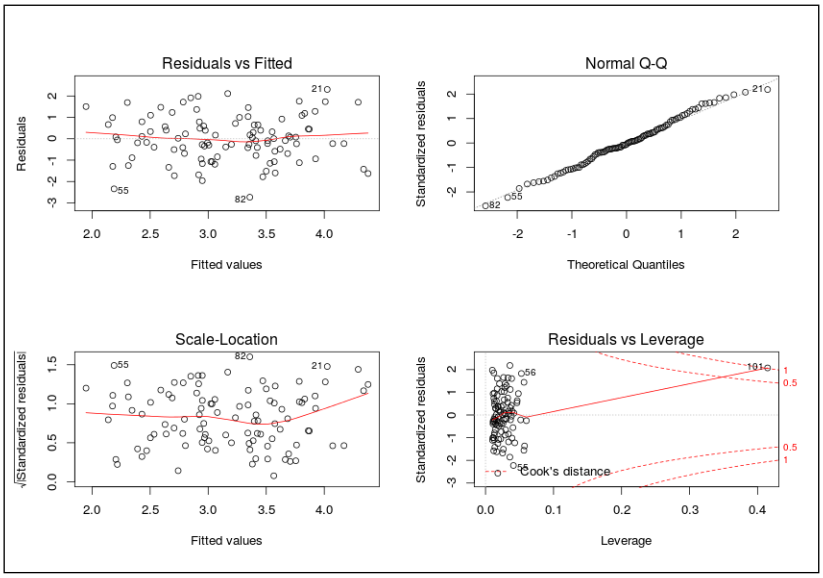
c. The p-values are 7.61e-15, 0.0487, 0.3754 meaning β_0 and β_1 are significant.
$\hat{\beta}_0 = 2.1305$
$\hat{\beta}_1 = 1.4396$
$\hat{\beta}_2 = 1.0097$
$\hat{\beta}_0$ appears close to the true value, however $\hat{\beta}_1$ and $\hat{\beta}_2$ are far from their true values.

d. Here the p-values are 8.27e-15 and 2.66e-06 meaning we can reject the null for both β_0 and β_1. The parameter estimates are fairly close to the true values. $\hat{\beta}_0 = 2.1124$
$\hat{\beta}_1 = 1.9759$

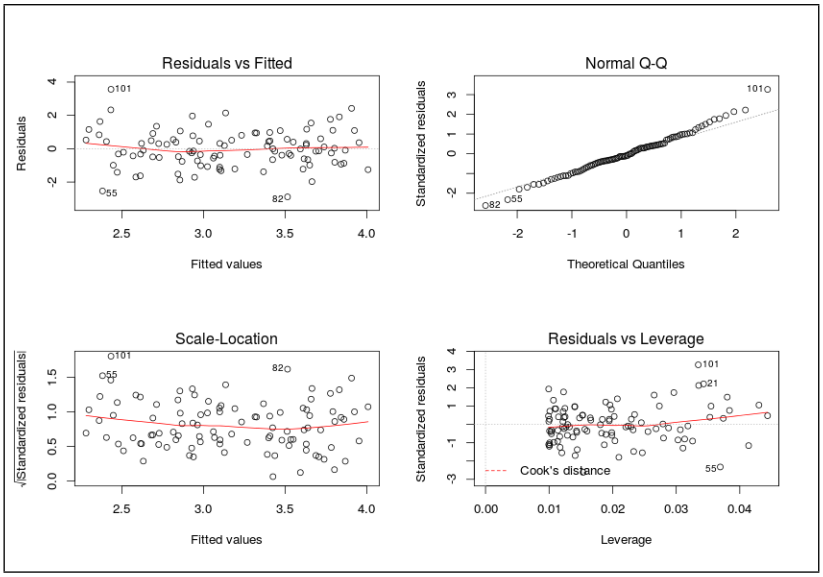
e. Here the p-values are $<$ 2e-16 and 1.37e-05 meaning we can reject the null for both β_0 and β_2. $\hat{\beta}_0 = 2.3899$
$\hat{\beta}_2 = 2.8996$

f. The results of (c)-(e) do not contradict each other. The data has been generated in such a way that x2(x1) meaning that we would expect the two variables to have a high degree of collinearity.

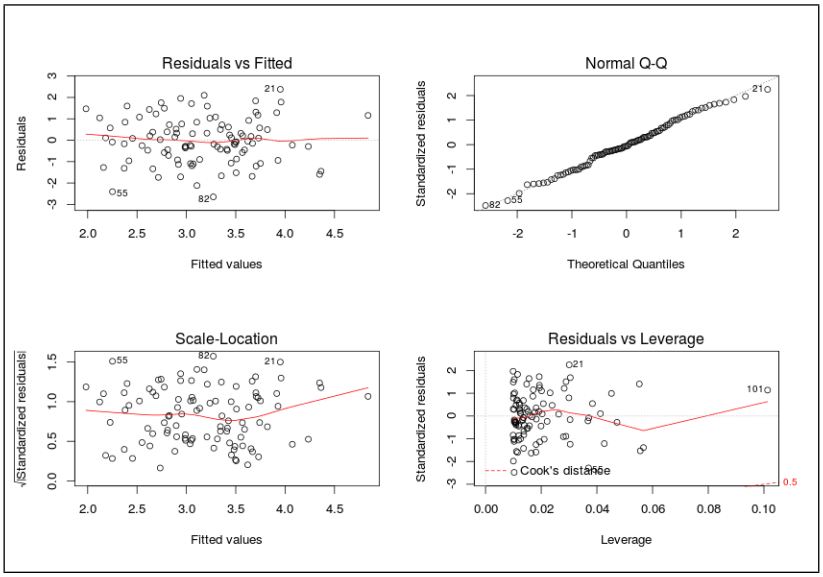
g. c. The p-values are 7.91e-16, 0.36458, and 0.00614 meaning β_0 and β_2 are significant. Here the new data point has made β_1 non-significant and β_2 significant. While there don't appear to be any outliers, point 101 is clearly a high leverage point.
$\hat{\beta}_0 = 2.2267$
$\hat{\beta}_1 = 0.5394$
$\hat{\beta}_2 = 2.5146$



d. The p-values are 1.78e-15 and 4.29e-05 meaning β_0 and β_1 are significant; the same as without the outlier. However, the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ have changed. In this case, point 101 appears to be an outlier and potentially a point of high leverage (depending on how we classify these points).
$\hat{\beta}_0 = 2.2569$
$\hat{\beta}_1 = 1.7657$



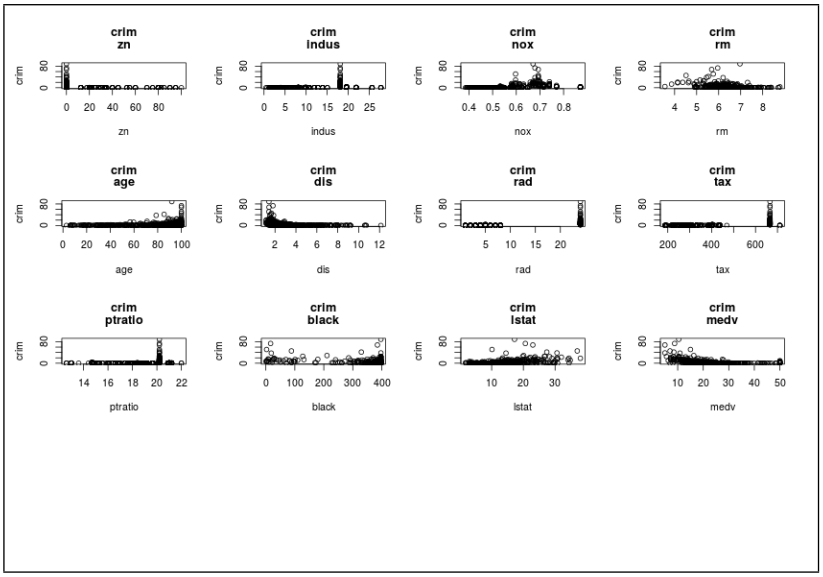
e. The p-values are $< 2e\text{-}16$ and 1.25e-06 meaning β_0 and β_2 are significant; the same as without the outlier. The values for $\hat{\beta}_0$ and $\hat{\beta}_2$ have changed. In this case, point 101 appears to be a high leverage point, but not an outlier.
$\hat{\beta}_0 = 2.3451$
$\hat{\beta}_1 = 3.1190$



Overall, inclusion of point 101 impacts all models and appears to be an outlier, a high leverage point, or both in all cases.

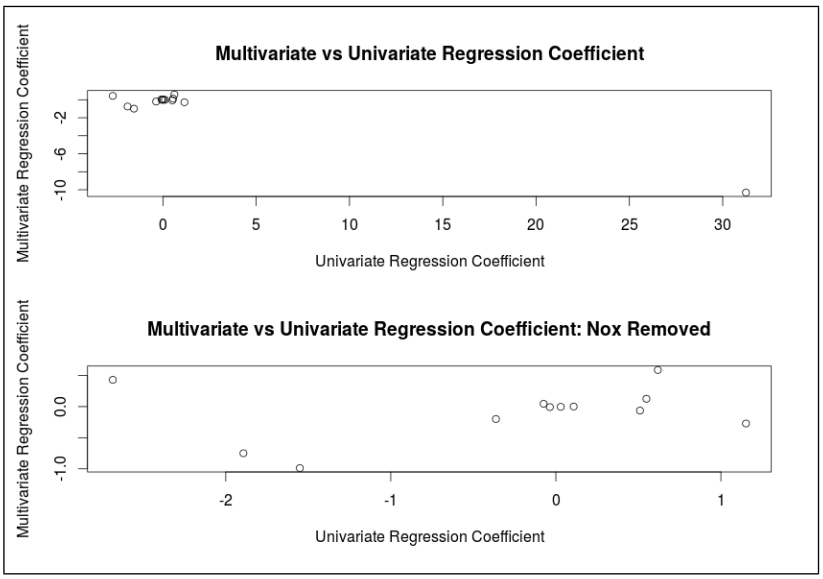
3.7.15

a. If we try to find significance using one predictor at a time, all predictors with the exception of chas have statistically significant relationships.



b. Here we can reject the null for the coefficients of 5 predictors: zn, dis, rad, black and medv. The p-value for the F-test is $< 2.2\text{e-}16$ and R^2 is 0.454.

c. All predictors with the exception of chas were significant in the first case while only zn, dis, rad, black and medv are significant here. This highlights the problem of attempting to find significant variables by performing individual regressions.



d. Using a cubic model, the following predictors have statistically significant coefficients for either X^2, X^3, or both: indus, nox, age, dis, ptratio and medv.

```
#############################################################################################
#################################### Supporting Code ########################################
#############################################################################################
# STAT4702 Homework 8
# Patrick Rogan
# UNI: psr2125
# 3.7.8 ------------------------------------------------------------------
Auto = read.csv("/home/pat/Desktop/STAT4702/ISL_Datasets/Auto.csv",
                header = T, na.strings = "?")
Auto = na.omit(Auto) #looking at the data, this does not have an impact on
#this exercise
attach(Auto)

mpgVSHP = lm(mpg ~ horsepower, data = Auto)

summary(mpgVSHP)

summary(mpgVSHP)$sigma/mean(mpg) # part ii
summary(mpgVSHP)$r.squared

predict(mpgVSHP, data.frame(horsepower = (c(98))), interval = "confidence")

plot(horsepower, mpg,main="MPG vs Horsepower",xlab = "Horsepower", ylab = "MPG")
abline(mpgVSHP)
par(mfrow = c(2,2))

plot(mpgVSHP)

# 3.7.9 ------------------------------------------------------------------
pairs(Auto)
options(digits = 4)
cor(Auto[,-9])

multiMPG = lm(mpg ~ cylinders + displacement + horsepower + weight +
                acceleration + year + origin, data = Auto)

summary(multiMPG)

multiMPG2 = lm(mpg ~ cylinders + displacement + horsepower + weight +
                acceleration + origin, data = Auto)

summary(multiMPG2)

plot(multiMPG)

mMPGwIntAll = lm(mpg ~ cylinders*displacement*horsepower*weight*
                acceleration*year*origin, data = Auto)

summary(mMPGwInt)

cor(data.frame(Auto$cylinders,Auto$displacement,Auto$horsepower,Auto$weight,
               Auto$acceleration,Auto$year,Auto$origin))

mMPGwInt2 = lm(mpg ~ cylinders+displacement+horsepower*weight+
                 displacement:acceleration+acceleration+year+origin, data = Auto)

summary(mMPGwInt2)

plot(mMPGwInt2)

#Non-linear effects
nlMPG = lm(mpg ~ cylinders + displacement + log(displacement) + horsepower +log(horsepower)  +
             weight + log(weight) + acceleration + year +I(year^2) + origin,data = Auto)
```

```
summary(nlMPG)
plot(nlMPG)

# 3.7.10 ------------------------------------------------------------------
rm(list = ls())
library(ISLR)
attach(Carseats)

mLMC = lm(Sales ~ Price + Urban + US, data = Carseats)
summary(mLMC)

mLMC2 = lm(Sales ~ Price + US, data = Carseats)
summary(mLMC2)

contrasts(US)

confint(mLMC2)

par(mfrow = c(2,2))
plot(mLMC2)
(2 + 1)/length(US)

# 3.7.11 ------------------------------------------------------------------
set.seed(1)
x = rnorm(100)
y = 2*x + rnorm(100)
noI = lm(y~x+0)
summary(noI)

noII = lm(x~y+0)
summary(noII)

tstat = sqrt(length(x)-1)*x%*%y/
  sqrt(x%*%x*y%*%y-(x%*%y)^2)

IM = lm(x~y)
IMI = lm(y~x)
summary(IM)$coefficients[6] - summary(IMI)$coefficients[6] #the two are

#equivalent down to the available precision
# 3.7.13 ------------------------------------------------------------------
set.seed(1)
X = rnorm(100) # N(0,1) default
eps = rnorm(100, mean = 0, sd = 0.25)

Y = -1 + 0.5*X + eps
length(Y)

plot(X,Y,main="Y vs X")

regYX = lm(Y ~ X)
summary(regYX)

abline(regYX,col="black",lwd=c(2.5,2.5))
abline(-1,0.5,col = "cyan",lwd=c(2.5,2.5))
leg.txt = c("Least Squares Line","Population Line")
legend(-2,y=0,leg.txt,
  lty=c(1,1),
  lwd=c(2.5,2.5),
  col=c("black","cyan"))

quadXY = lm(Y ~ X + I(X^2))

summary(quadXY)
```

```
confint(regYX)

# h
set.seed(1)
X = rnorm(100) # N(0,1) default
eps = rnorm(100, mean = 0, sd = 0.05)

Y = -1 + 0.5*X + eps
length(Y)

plot(X,Y,main="Y vs X")

regYX = lm(Y ~ X)
summary(regYX)

abline(regYX,col="black",lwd=c(2.5,2.5))
abline(-1,0.5,col = "cyan",lwd=c(2.5,2.5))
leg.txt = c("Least Squares Line","Population Line")
legend(-2,y=0,leg.txt,
       lty=c(1,1),
       lwd=c(2.5,2.5),
       col=c("black","cyan"))

confint(regYX)

# i
set.seed(1)
X = rnorm(100) # N(0,1) default
eps = rnorm(100, mean = 0, sd = 2)

Y = -1 + 0.5*X + eps
length(Y)

plot(X,Y,main="Y vs X")

regYX = lm(Y ~ X)
summary(regYX)

abline(regYX,col="black",lwd=c(2.5,2.5))
abline(-1,0.5,col = "cyan",lwd=c(2.5,2.5))
leg.txt = c("Least Squares Line","Population Line")
legend(-2,y=4,leg.txt,
       lty=c(1,1),
       lwd=c(2.5,2.5),
       col=c("black","cyan"))
confint(regYX)

# 3.7.14 ------------------------------------------------------------------
set.seed(1)
x1 = runif(100)
x2 =0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)

plot(x1,x2, main="x2 vs x1")

regYX1X2 = lm(y~x1+x2)

summary(regYX1X2)

regYX1 = lm(y~x1)

summary(regYX1)
```

```
regYX2 = lm(y~x2)
summary(regYX2)

x1 = c(x1, 0.1)
x2 = c(x2, 0.8)
y = c(y, 6)

regYX1X2 = lm(y~x1+x2)
summary(regYX1X2)
par(mfrow = c(2,2))
plot(regYX1X2)

regYX1 = lm(y~x1)
summary(regYX1)
par(mfrow = c(2,2))
plot(regYX1)

regYX2 = lm(y~x2)
summary(regYX2)
par(mfrow = c(2,2))
plot(regYX2)

# 3.7.15 ------------------------------------------------------------------
library(MASS)
dim(Boston)
par(mfrow = c(4,4))

linearPredictors = c(1:(length(colnames(Boston))-1))

for (x in 2:length(colnames(Boston))){
  cm = lm(Boston$crim ~ Boston[,x])
  sm = summary(cm)
  print(summary(cm))
  linearPredictors[(x-1)] = sm$coefficients[2]
  if (sm$coefficients[8] < 0.05){
    cat(colnames(Boston[x]),"p-value", sm$coefficients[8],"\n")
    plot(Boston[,x], Boston[,1], xlab=colnames(Boston[x]),
         ylab = "crim",main=c("crim vs",colnames(Boston[x])))
  }
}

multiPredictors = c(1:(length(colnames(Boston))-1))

attach(Boston)
multiB = lm(crim ~. , data = Boston)
smM = summary(multiB)
multiPredictors = smM$coefficients[2:14]
par(mfrow=c(2,1))
plot(linearPredictors,multiPredictors,xlab = "Univariate Regression Coefficient",
     ylab = "Multivariate Regression Coefficient",
     main="Multivariate vs Univariate Regression Coefficient")

plot(linearPredictors[-4],multiPredictors[-4],xlab = "Univariate Regression Coefficient",
     ylab = "Multivariate Regression Coefficient",
     main="Multivariate vs Univariate Regression Coefficient: Nox Removed")


for (x in 2:length(colnames(Boston))){
  cm = lm(Boston$crim ~ Boston[,x] + I(Boston[,x]^2) + I(Boston[,x]^3))
  sm = summary(cm)
  print(colnames(Boston)[x])
  print(summary(cm))
}
```