Homework 2

Hyeonjin Lee

2/27/2021

Introduction to Statistical Learning in R Homework 2 (Chapter 3):

Question 4

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression

a) Suppose that the true relationship between X and Y is linear. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

There isn't complete concrete information to fully determine which training RSS is lower betwee linear and cubic. However, because the true relationship is determined to be linear, we would expect the least squares line to be similar to the true regression line and thus the RSS for the linear regression may be lower than the cubic.

b) Answer (a) using test rather than training RSS.

There isn't enough information as the test RSS depends on the test data. However, the cubic regression will have a higher test RSS due to the overfitting that which may have more error than the linear regression.

c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Cubic regression will have a lower train RSS because of the higher flexibility. Therefore, it will flow the points of the training model more closely and lower the train RSS.

d) Answer (c) using test rather than training RSS.

When referring to the test RSS, there is not enough information to conclude which test RSS will be lower. The claim is also that we don't know how far it is from linear. If it was closer to the linear than cubic, the test RSS for linear will be smaller and vice versa.

Question 10

This question should be answered using the Carseats data set.

a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```r
library(ISLR)
data(Carseats)
fit1 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit1)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative

Because sales is in units in thousands, we can multiply the coefficients by 1000 to get one unit in sales.

On average, for every dollar increase in Price, we can expect a 54.46 unit decrease in sales considering all other predictors are fixed.

On average, the unit sales in Urban locations are 21.916 units less than rural locations.

On average, the unit sales in the US locations are 1200.57 units greater than non-US locations.

c) Write out the model in equation form, being careful to handle the qualitative variables properly.

Sales=13.0434689 + (-0.0544588)Price + (-0.0219162)Urban + (1.2005727)US + ϵ
Urban = 1 if the store is in an Urban location and 0 otherwise
US = 1 if the store is in a US location and 0 otherwise

d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

We can see that the p-value of Price and US variables are extremely small and smaller than our alpha of .05. Therefore, we can reject the null hypothesis for those variables.

e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
fit2 <- lm(Sales ~ Price + US, data = Carseats)
summary(fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f) How well do the models in (a) and (e) fit the data?

In both models, the R-squared value is .2393. This shows that only 23.93% of the variability is explained by the model. The R-squared for the smaller model is very similar if not slightly better than the bigger model in part a.

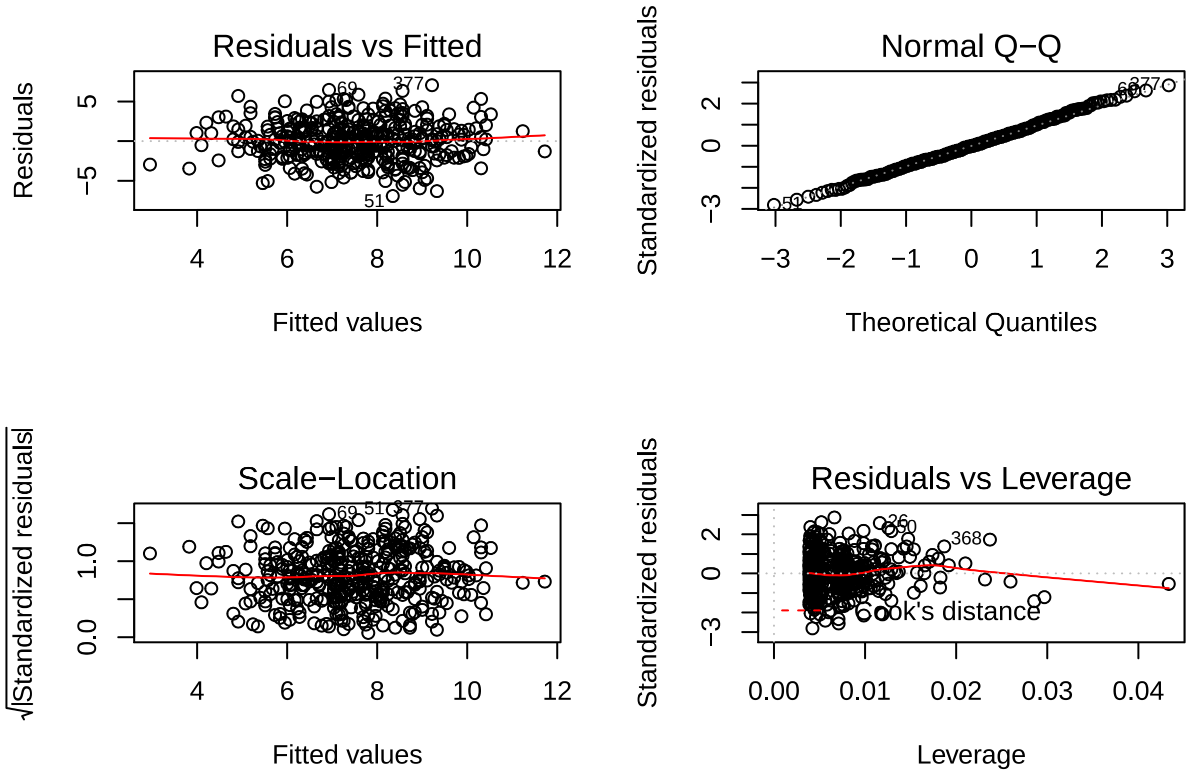
g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

```
confint(fit2)
```

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

h) Is there evidence of outliers or high leverage observations in the model from (e)?

```
par(mfrow=c(2,2))
plot(fit2)
```



> Looking at the plot of residuals vs leverage, we can see evidence of a few outliers as some values are higher than 2 and lower than -2. There is also evidence of some leverage points as points exeed (p + 1)/ n(.01).

Question 13

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

a) Using the rnorm() function, create a vector, x, containing 100 observations drawn from a N(0, 1) distribution. This represents a feature, X.

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

b) Using the rnorm() function, create a vector, eps, containing 100 observations drawn from a N(0, 0.25) distribution i.e. a normal distribution with mean zero and variance 0.25.

```
eps <- rnorm(100, sd = sqrt(.25))
```

c) Using x and eps, generate a vector y according to the model Y = -1+0.5X + ϵ
What is the length of the vector y? What are the values of $\beta 0$ and $\beta 1$ in this linear model?

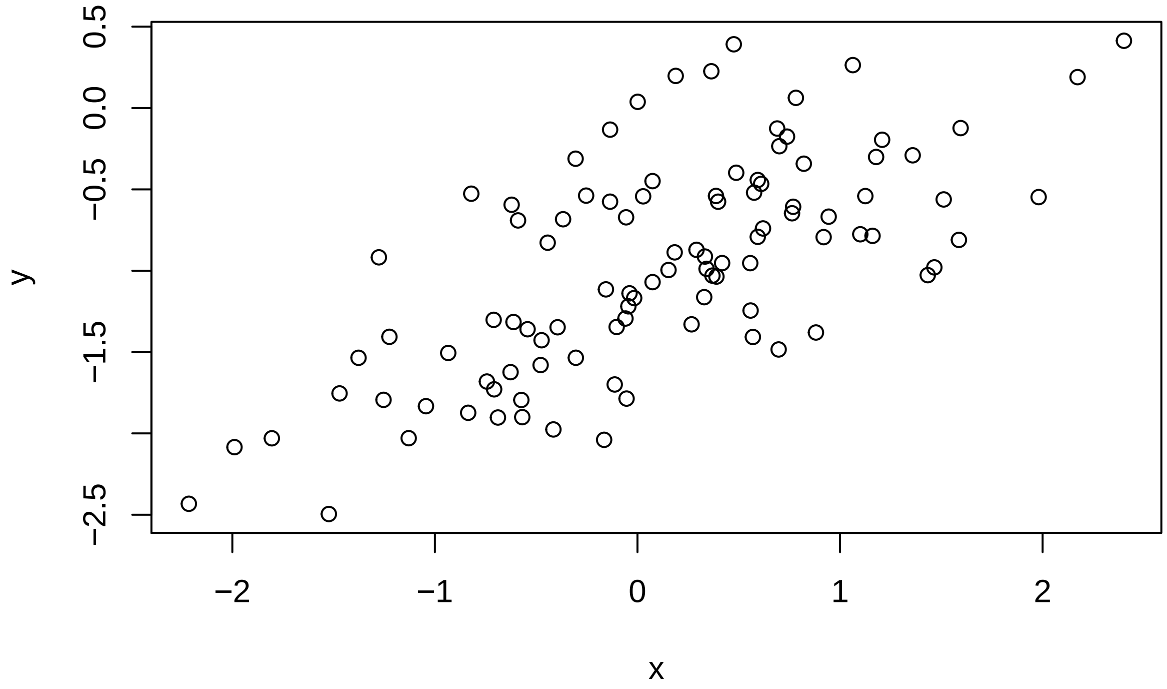
```
y <- -1 + .5 * x + eps
length(y)
```

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

values of $\beta 0$ and $\beta 1$ are -1 and .5

d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

```
plot(x,y)
```



> For the most part, x and y look to have a linear relationship with some variance from the eps.

e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta}0$ and $\hat{\beta}1$ compare to $\beta 0$ and $\beta 1$?

```
fit3 <- lm(y ~ x)
summary(fit3)
```

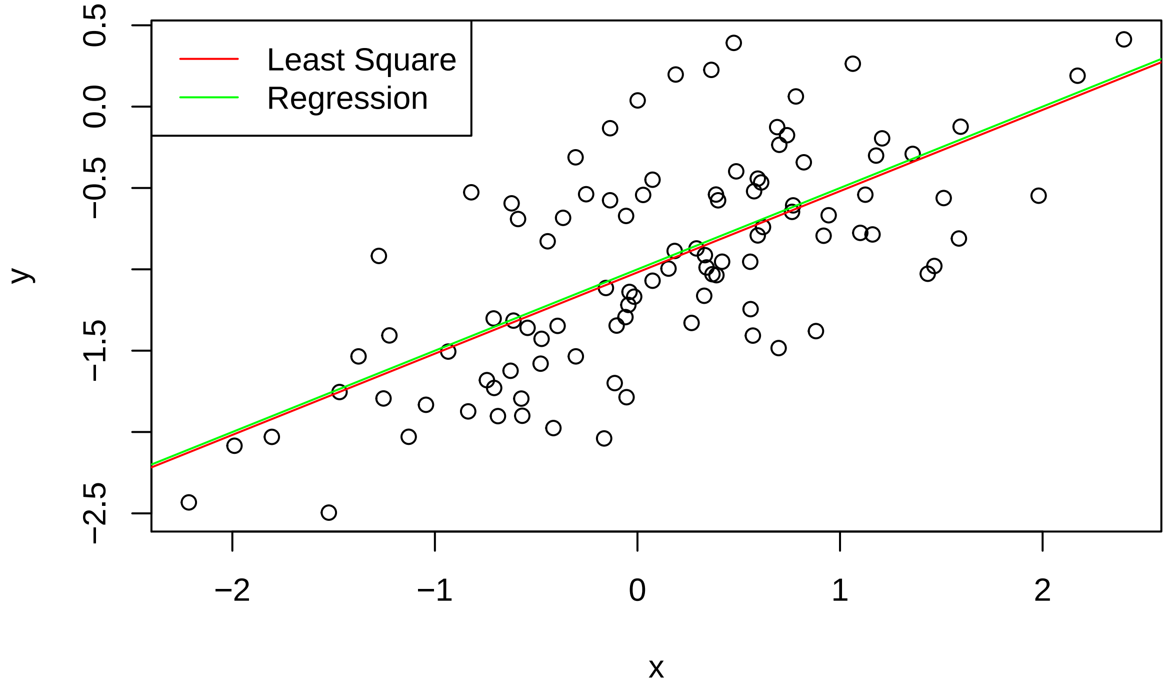
```
##
## Call:
## lm(formula = y ~ x)
##
```

```
## 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 model has a p-value that is extremely small and smaller than our alpha of .05 therefore we can reject the null hypothesis. In addition, $\hat{\beta}0$ and $\hat{\beta}1$ are -1.01885 and .49947 respectively. This is very close to the $\beta0$ and $\beta1$ values of -1 and .5.

f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.

```
plot(x,y)
abline(fit3, col="red")
abline(-1,.5, col="green")
legend("topleft", c("Least Square", "Regression"), col = c("red", "green"),lty=c(1,1))
```



g) Now fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

```
fit4 <- lm(y ~ x + I(x^2))
summary(fit4)
```

```
##
## 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
```

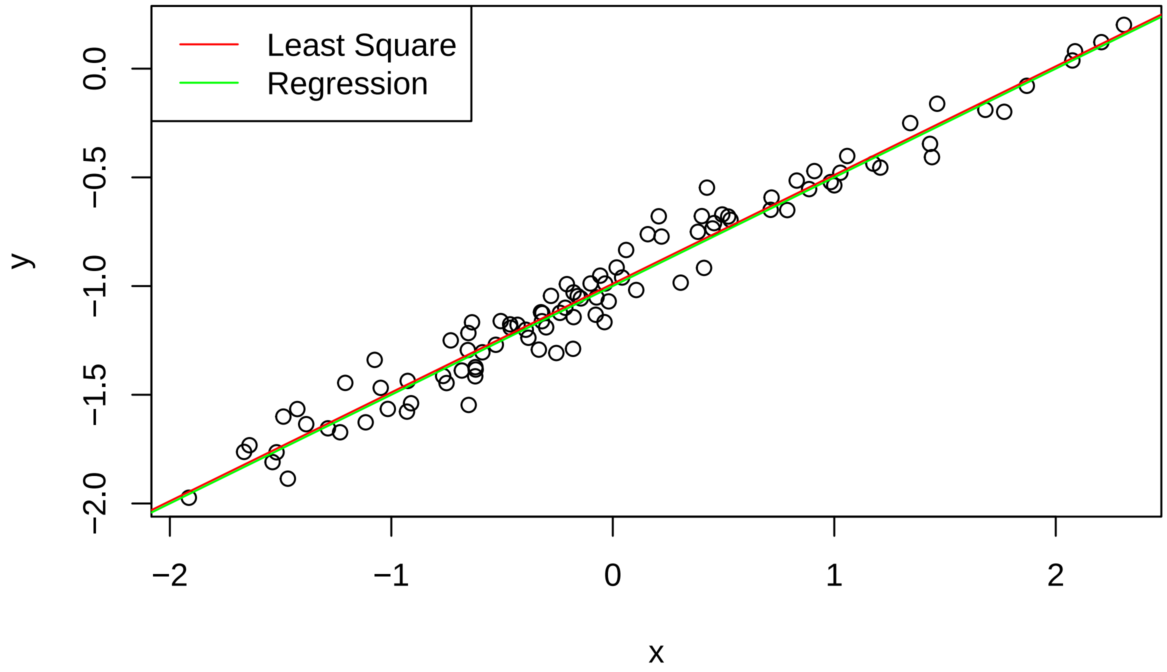
Although there was an increase in the R-squared value and a decrease in the RSE, we can note that (xˆ2) term is not significant. The p-value is greater than .05 and thus there is not enough evidence to suggest that the new model improves the model fit.

h) Repeat (a)–(f) after modifying the data generation process in such a way that there is less noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

```
set.seed(1)
eps <- rnorm(100, sd= .1)
x <- rnorm(100)
y <- -1 + .5*x + eps
plot(x,y)
fit5 <- lm(y ~ x)
summary(fit5)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.232416 -0.060361  0.000536  0.058305  0.229316
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.989115   0.009035 -109.48   <2e-16 ***
## x            0.499907   0.009472   52.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09028 on 98 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9657
## F-statistic:  2785 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
abline(fit5, col="red")
abline(-1,.5, col="green")
legend("topleft", c("Least Square", "Regression"),col = c("red", "green"), lty=c(1,1))
```



> By decreasing the variance of the normal distribution of the error term, the values follow a more strictly linear pattern. This has led to a decrease in RSE and an increase in Rˆ2. The two lines also overlap each other as the slope and intercept are nearly the same.

i) Repeat (a)–(f) after modifying the data generation process in such a way that there is more noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

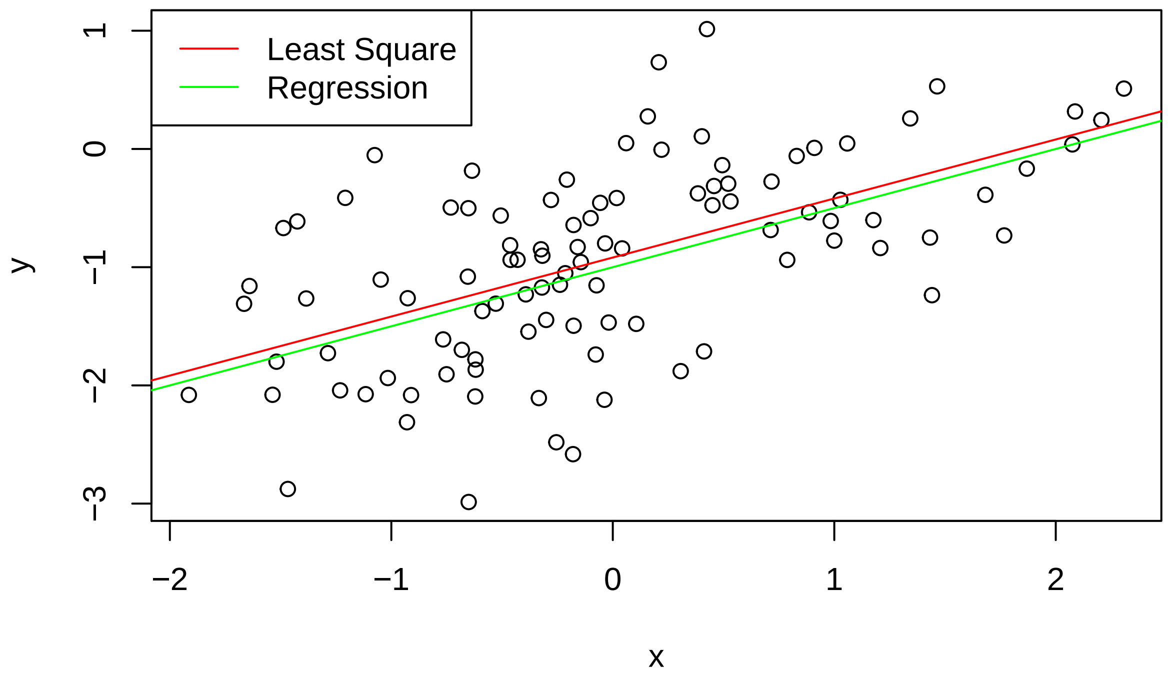
```
set.seed(1)
eps <- rnorm(100, sd=.75)
x <- rnorm(100)
y <- -1 + .5 * x + eps
plot(x,y)

fit6 <- lm(y~x)
summary(fit6)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74312 -0.45271  0.00402  0.43728  1.71987
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.91836    0.06776 -13.553  < 2e-16 ***
## x            0.49930    0.07104   7.028 2.81e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6771 on 98 degrees of freedom
## Multiple R-squared:  0.3351, Adjusted R-squared:  0.3283
## F-statistic:  49.4 on 1 and 98 DF,  p-value: 2.808e-10
```

```
abline(fit6, col="red")
abline(-1,.5, col="green")
legend("topleft", c("Least Square", "Regression"), col=c("red", "green"), lty=c(1,1))
```



> We increased the variance of the error term drastically. We can see that the linear relationship has become weaker and thus a lower Rˆ2 and a higher RSE. In addition, the lines do not overlap as close as the other models but they are still fairly similar as seen by the intercept and slope values.

j) What are the confidence intervals for $\beta0$ and $\beta1$ based on the original data set, the noisier data set, and the less noisy data set? Comment on your results

```
# Original Data
confint(fit3)
```

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

```
#Less Noisy Data
confint(fit5)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0070441 -0.9711855
## x            0.4811096  0.5187039
```

```
# Noisier Data
confint(fit6)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0528304 -0.7838914
## x            0.3583218  0.6402796
```

We can note that as the noise increased, the confidence interval became wider while as the noise decreased, the confidence interval became narrower. In addition, all intervals seem to be centered around .5 which is appropriate.

Question 14 (a-f)

This problem focuses on the collinearity problem

a) Perform the following commands in R:

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

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

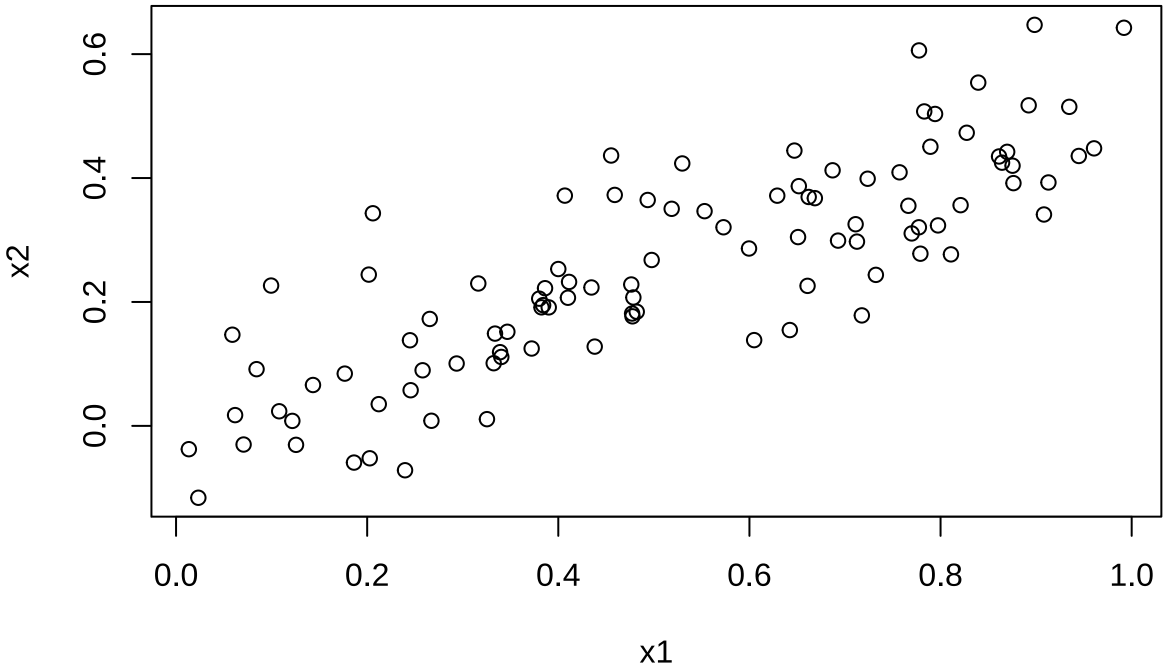
Y = 2 + 2 X1 + .3X2 + ϵ The regression coefficients are 2, 2, and .3 respectively.

b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

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

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

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



c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\hat{\beta}0$, $\hat{\beta}1$, and $\hat{\beta}2$?

How do these relate to the true $\beta0$, $\beta1$, and $\beta2$?

Can you reject the null hypothesis $H_0 : \beta1 = 0$? How about the null hypothesis $H_0 : \beta2 = 0$?

```r
fit7 <- lm(y ~ x1 + x2)
summary(fit7)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

$\hat{\beta}0$, $\hat{\beta}1$, and $\hat{\beta}2$ are 2.13, 1.4396, and 1.0097. Only $\hat{\beta}0$ is close to $\beta0$ of 2. We also can see that the p-value of x2 is greater than .05 and thus cannot reject the null hypothesis for $\beta2$ We can reject the null hypothesis on $\beta1$ as the p-value is less than .05 and thus significant.

d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta1 = 0$

```r
fit8 <- lm(y ~ x1)
summary(fit8)
```

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

We can reject the null hypothesis $H_0 : \beta1 = 0$ as the p-value is extremely small and smaller than our alpha of .05. We can also note that the coefficient for x1 is much higher in the newer model than the previous.

e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta1 = 0$

```r
fit9 <- lm(y ~ x2)
summary(fit9)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

$H_0 : \beta 1 = 0$ can be rejected as the p-value is extremely small and smaller than our alpha of .05 and thus significant. We can also see that the coefficient for x2 is much higher in this new model.

f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

There seems to be evidence of collinearity between the predictors x1 and x2. We can see when comparing the individual models that the significance of the x2 variable is hidden due to the collinearity between the variables. We can also see that the standard error for $\hat{\beta}1$ is greater with the model that includes both x1 and x2 and we should note that the collinearity reduces the accuracy of the estimates of the regression coefficients.