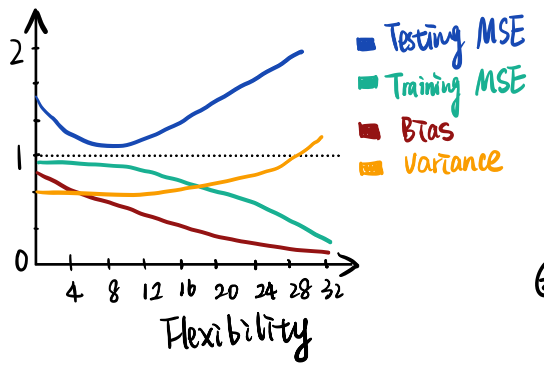
STA325 HW1

Kedi (Sophia) Li
2022/09/11

Question 1

a. The model fits the data well, Because the training data fall above and below the non-linear regression model so that errors have approximately zero mean. However, somewhere around $x = 14$, the model underestimate and around $x = 18$ to $x = 22$, the model overestimate.

b.



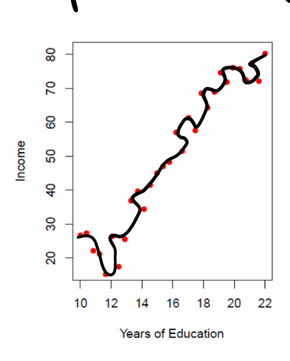
① The testing MSE will be higher than training MSE because of overfitting with high flexibility

② The Bias will decrease as we fit with a more dots-connected model that reduces the noise term ε to approach zero.

③ The variance will increase because we contain more parameters in our more-flexible models.

Any model with high flexibility will fit the model with lower bias but become less interpretable than simple linear model with higher bias but lower variance of model misspecification.
Also, the more flexible model will run into the problem of overfitting and we should take into account the trade off between variance and bias. therefore, the variance will decrease first and increase afterwards when the model becomes more flexible

c. If we assume high-flexibility model by using the "connect-the-dots" method



the fitted model $\hat{f}(\cdot)$ will be like:

① The training MSE $= Ave_{i \in Tr}[y_i - \hat{f}(x_i)]^2$, which the bias ter will be minized to nearly zero, traing MSE also decreases.

② It falls into overfitting issue and becomes less interpretable with high variance. When we draw a new data set from distribution, this model $\hat{f}(\cdot)$ will not perform well on new data set, the testing MSE will increase afterwards.

d. The income around $x = 12$ fluctuates or even decreases, but it increases a lot after $x = 12$ and keeps a steeper slope at $x = 12$, indicating the truth that having only high-school degree might not help with income, but having a bachelor-degree or any diploma higher than high-school is associated with higher income. Thus, I will suggest students to stick on their study and complete bachelor-degree in 4 yrs.

Question 2

a. $Y(x)$ is random because $Y(x) = f(x) + \varepsilon$ contains the noise term ε.
$\hat{C}(x)$ is not random because the input value x has been fixed.
X is not random

b. The lowest test error is given by the Bayes classifier, which is the Bayes error rate, written as $1 - E(\max_k Pr(Y=k|X))$, which is to minimize $1 - P_k(x)$, we will like to find $g(x) = \hat{C}(x) = k^*$

therefore, if we want this error rate to be minimized, we'd like to find the k^* category that generates the biggest probability with x, where $k^* = \arg\max p_k(x)$ for $k = 1, \cdots, K$; thus, the $MCE(x)$ will be the smallest.

c. The point of the classifier is to present each individual observation to the most accurate and most likely category by its predictor values.

thus, we should assign an individual observation with a predictor X_i for $i = 1, \cdots, n$ to the category k where $P(Y=k|X=x_i)$ is the largest one, for which the error of miss-classifying is minimized, thus, we choose Bayer optimal classifier

d. In reality, we don't know the probability for each of the observation for each of the category k, therefore we actually use estimated probability when we determine the optimal classifer, leading to the fact that true classifier will never be the same with the Bayer Optimal one.

Question 3

a) β_0 is the intercept of regression meaning the expected number of hours spent by freshman.

b) $(\beta_0 + \beta_1)$ is the expected number of hours spent by seniors.

c) β_1 is the expected extra number of hours spent by seniors over freshmen.

d) The R^2 is the fraction of variability in number of hours that can be explained by shool year, while this model capture only the class status of students, which many omitted many other variable such as study habit, major, and any confounding variables therefore, this simple linear model will have small R^2 and should have counted more β.

Question 4

a) $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$

$\left(\frac{p(x)}{1-p(x)}\right) = e^{\beta_0+\beta_1 x}$ ↗

$p(x) = (e^{\beta_0+\beta_1 x})(1-p(x))$

$p(x) = e^{\beta_0+\beta_1 x} - p(x)\cdot e^{\beta_0+\beta_1 x}$

$p(x) = e^{\beta_0+\beta_1 x} / 1+e^{\beta_0+\beta_1(x)}$

b) There is a non-linear relationship between x and $p(x)$

c) The log odd-ratio is expected to increase by β_1 with one unit increase in the predictor x.

Question 7:

The logistic regression perfoms better:

1) The 1-nearest neighbors is much more flexible with the issue of overfitting, If we draw a new data set from the distribution, the 1-nearest neighbors will not perform well and have little inference values.

2) The logistic regression is better in case of its lower test error rate.

For the use of 1-nearest neighbors that KNN with K=1, we have a training error of 0% because $P(Y=j|X=x_i) = I(y_i=j)$, where will equal to 1 if $y_i=j$ and 0 if not.

Thus, there is no error made on the training data, but we have an average error rate of 18% indicating that the test error rate will double if KNN, comparing with the logistic regression model with test error rate of 30%, the 1-nearest neighbor method perform worse.

Question 8:

a. $\hat{f}(X)$ is not random, ε is random

$Y(x)$ is random, x is not random.

b. Z is a continous random variable with distribution function $F(\cdot)$

$dF(z) = f(z)dz \quad E|Z-M| = \int_{-\infty}^{z=m} (m-z)f(z)dz + \int_{z=m}^{\infty} (z-m)f(z)\,dz$

STA325_HW1_KediLi

R Markdown

This question involves the use of multiple linear regression on the Auto data set.

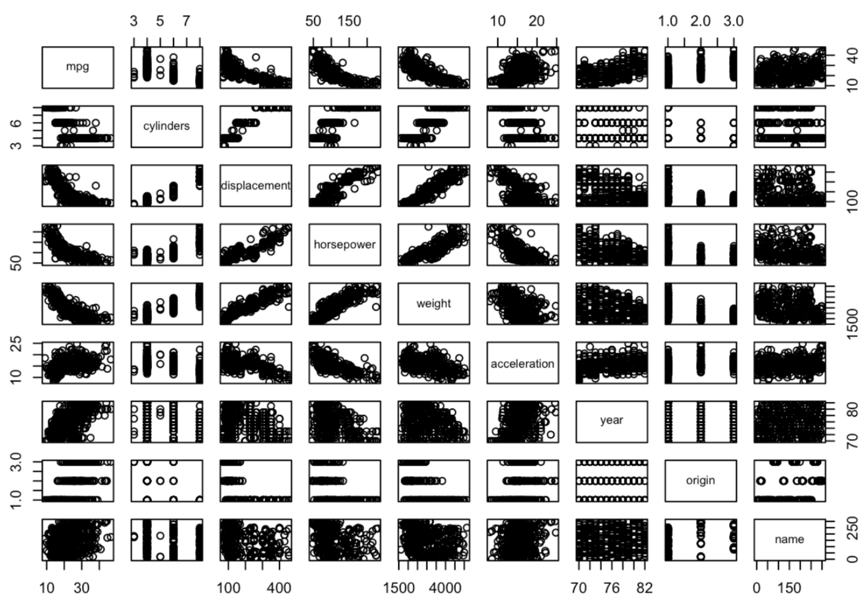
##QUESTION 5: (a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
data("Auto")
Auto
```

	mpg <dbl>	cylinders <dbl>	displacement <dbl>	horsepower <dbl>	weight <dbl>	acceleration <dbl>	year <dbl>	origin <dbl>
1	18.0	8	307.0	130	3504	12.0	70	1
2	15.0	8	350.0	165	3693	11.5	70	1
3	18.0	8	318.0	150	3436	11.0	70	1
4	16.0	8	304.0	150	3433	12.0	70	1
5	17.0	8	302.0	140	3449	10.5	70	1
6	15.0	8	429.0	198	4341	10.0	70	1
7	14.0	8	454.0	220	4354	9.0	70	1
8	14.0	8	440.0	215	4312	8.5	70	1
9	14.0	8	455.0	225	4425	10.0	70	1
10	15.0	8	390.0	190	3850	8.5	70	1

1-10 of 392 rows | 1-9 of 10 columns Previous **1** 2 3 4 5 6 ... 40 Next

```
pairs(Auto)
```



From this pair-wise draft, it is clear that there are several moderately positive/negative relationship between predictors; for example, the weight and mpd displays a negative association as higher weight indicates lower mpg, and there is a positive relationship between horsepower and displacement.

b. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

```
cor(Auto[,-9])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

The matrix of correlations between variables indicate the possible collinearity between predictors for which those have a correlation bigger than 0.5: the displacement and cylinder(0.95), the acceleration and horsepower(0.68), etc. When we plot the multi-linear regression model, we should take interactions between those predictors into consideration in order to better fit the model.

c. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

```
multi_linear_model <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin, dat
a = Auto)
summary(multi_linear_model)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

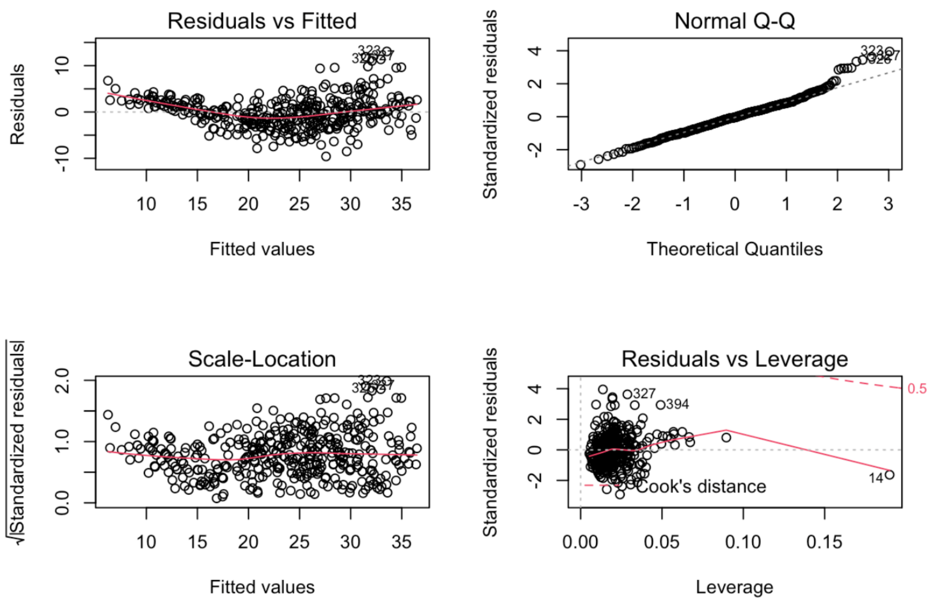
i. There is indeed a relationship between predictors and the response variable mpg as the overall p-value is smaller than 0.05 indicating a strong association between predictors and response variable, and the adjusted r-squared of 0.8182 means that about 81.82% of variability was captured in the model and explained by those variables, which is a optimal model as a function of mpg.

ii:To look at the P-value, the predictors displacement, weight, year, and origin have strong relationships with the response variable mpg because they all contain small p-value that smaller than 0.05 to reject the null hypothesis, meaning that there are statistically significant.

iii: The coefficient for year variable is 0.751, indicating that for one unit increase in the year is associated with 0.751 unit expected increase in mpg generally, with holding all other variables unchanged.

d. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

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




```
leverage = 2*(1+7)/nrow(Auto)
leverage
```

```
## [1] 0.04081633
```

To check to assumptions we will look at the shape, constant variance, normality, independence, and randomness assumption. For the shape and the constant variance assumption, the residual plot shows a cone shape pattern as the residuals are getting larger with the increase of the fitted value, therefore, the constant variance assumption is not valid and indicating a somehow questioned linear model. The QQ-plot validate the normality of the regression model as the residuals appear linear but only skewed a little at the right tail, which, for a large amount of data points, is normal. Furthermore, the Scale-Location plot shows no outliers because no residuals located outside of two standard deviations range. And the leverage is about 0.041, it is apparently some of the residuals located exceeding 0.041 threshold, but the cook's distance was not been violated meanings that no influential outliers existed for the regression model.

e. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
new_multi_linear_model <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin
+ horsepower*acceleration + cylinders * displacement, data = Auto)
summary(new_multi_linear_model)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin + horsepower * acceleration +
##     cylinders * displacement, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7832  -1.7034   0.0136   1.5496  12.0766
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.601e+01  5.538e+00  -2.890  0.00407 **
## cylinders               -1.816e+00  4.503e-01  -4.034 6.63e-05 ***
## displacement            -7.324e-02  1.388e-02  -5.277 2.21e-07 ***
## horsepower               5.356e-02  2.703e-02   1.982  0.04825 *
## weight                  -3.885e-03  6.882e-04  -5.645 3.23e-08 ***
## acceleration             6.543e-01  1.655e-01   3.954 9.17e-05 ***
## year                     7.608e-01  4.630e-02  16.432  < 2e-16 ***
## origin                   6.175e-01  2.686e-01   2.299  0.02203 *
## horsepower:acceleration -7.930e-03  1.854e-03  -4.277 2.40e-05 ***
## cylinders:displacement   1.050e-02  1.833e-03   5.727 2.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.022 on 382 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8501
## F-statistic: 247.4 on 9 and 382 DF,  p-value: < 2.2e-16
```

To deal with the multi-collinearity problem within the predictors, we can add interaction term to capture the association between predictors in the multi-linear regression model: thus, I add an interaction between horsepower and acceleration because I believe that these two predictors mutually affect each other and the reason that they are both not statistically significant is due to their internal-correlation. As the horsepower inform how much power the engine makes, and the acceleration is directly related to the capacity of engine. Similarly, I also add an interaction between cylinders and displacement because cylinder tells how big the engine size is and the displacement is a measure of cylinder volume. After adding two interactions, all predictors seem statistically significant as their p-value are all smaller than 0.05, and the adjusted r-squared increases from 0,81 to 0.85, indicating that the interactions between variables capture an extra of 4% variability in the model.

f. Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

i. Taking the log value of weight, all predictors are still statistically significant and the adjusted r-squared increase a little bit.

```
multi_linear_model1 <- lm(mpg ~ cylinders + displacement + horsepower + log(weight) + acceleration + year + origi
n + horsepower*acceleration + cylinders * displacement, data = Auto)
summary(multi_linear_model1)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + log(weight) +
##     acceleration + year + origin + horsepower * acceleration +
##     cylinders * displacement, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1057  -1.7017   0.0087   1.5452  12.1638
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              79.352586  14.350977   5.529 5.97e-08 ***
## cylinders                -1.166295   0.447581  -2.606 0.009525 **
## displacement             -0.043503   0.014849  -2.930 0.003596 **
## horsepower                0.060081   0.026052   2.306 0.021636 *
## log(weight)             -14.365724   1.961873  -7.322 1.45e-12 ***
## acceleration              0.651721   0.160600   4.058 6.00e-05 ***
## year                      0.781836   0.045375  17.231  < 2e-16 ***
## origin                    0.573551   0.261636   2.192 0.028970 *
## horsepower:acceleration  -0.007425   0.001762  -4.214 3.14e-05 ***
## cylinders:displacement    0.006546   0.001876   3.489 0.000541 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.945 on 382 degrees of freedom
## Multiple R-squared:  0.8609, Adjusted R-squared:  0.8576
## F-statistic: 262.6 on 9 and 382 DF,  p-value: < 2.2e-16
```

ii. Taking log value of horsepower, weight, acceleration and the square root of year, we generate a model with higher adjusted r-squared but those logged predictors perform a higher p-value indicating that they are not that significant comparing with their original values.

```
multi_linear_model2 <- lm(mpg ~ cylinders + displacement + log(horsepower) + log(weight) + log(acceleration) + sq
rt(year) + origin + log(horsepower)*log(acceleration) + cylinders*displacement, data = Auto)
summary(multi_linear_model2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + log(horsepower) +
##     log(weight) + log(acceleration) + sqrt(year) + origin + log(horsepower) *
##     log(acceleration) + cylinders * displacement, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9357 -1.5983  0.0205  1.5260 12.0859
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -47.738959  37.972421  -1.257  0.20945
## cylinders                          -1.101612   0.426000  -2.586  0.01008 *
## displacement                       -0.037537   0.014507  -2.587  0.01004 *
## log(horsepower)                    14.560217   7.351156   1.981  0.04835 *
## log(weight)                       -12.368897   2.037834  -6.070 3.09e-09 ***
## log(acceleration)                  33.581858  11.775062   2.852  0.00458 **
## sqrt(year)                         13.380652   0.789216  16.954  < 2e-16 ***
## origin                              0.618384   0.259552   2.383  0.01768 *
## log(horsepower):log(acceleration)  -8.019798   2.566755  -3.124  0.00192 **
## cylinders:displacement              0.005666   0.001797   3.153  0.00174 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.919 on 382 degrees of freedom
## Multiple R-squared:  0.8634, Adjusted R-squared:  0.8602
## F-statistic: 268.3 on 9 and 382 DF,  p-value: < 2.2e-16
```

QUESTION 6:

a. Perform the following commands in R: 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?

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

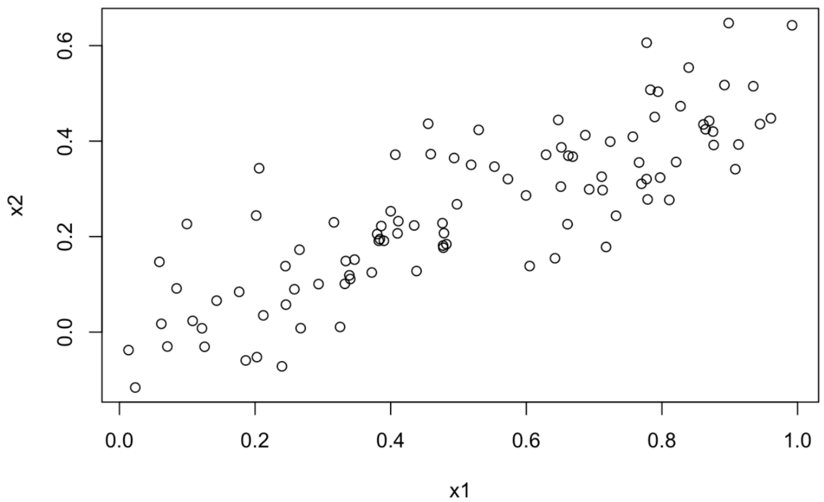
The linear model can be written as $y = 2 + 2X_1 + 3X_2 + \varepsilon$, for $\epsilon_i \overset{iid}{\sim} N(0, 1)$, and the $\beta_0 = 2$, $\beta_1 = 2$, and $\beta_2 = 0.3$.

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)
```



The scatterplot between x1 and x2 variables

The correlation of 0.84 indicates that the relationship between x1 and x2 is pretty strong. From the scatterplot, we can tell a relatively strong, positive, and linear relationship of x2 as a function of x1.

c. Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are βˆ0, βˆ1, and βˆ2? How do these relate to the true β0, β1, and β2? Can you reject the null hypothesis H0 : β1 = 0? How about the null hypothesis H0 : β2 = 0?

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

```
##
## 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
```

The linear model can be written as $\hat{y} = 2.13 + 1.44\hat{X_1} + 1.01\hat{X_2} + \epsilon$, for $\epsilon_i \overset{iid}{\sim} N(0,1)$, and the $\hat{\beta}_0 = 2.13$, $\hat{\beta}_1 = 1.44$, $\hat{\beta}_2 = 1.01$.Compare these estimated parameters with true parameters, $\beta_0 - \hat{\beta}_0 = -0.13$, $\beta_1 - \hat{\beta}_1 = 0.56$, and $\beta_0 - \hat{\beta}_0 = -0.71$, and we can see a moderate bias happened between estimated parameters and true parameters when building up a linear regression model. The $\hat{X_1}$ contains a p-value of 0.048 which is smaller than the threshold of 0.05, therefore we have evidence to say that the probability of observing such t-value of $\hat{X_1}$ is about 0.048 under the H0, so that we can reject the null hypothesis. However, the $\hat{X_2}$ contains a p-value of 0.38, which is much more bigger than 0.05, thus we don't have enough evidence to reject the null hypothesis because the probability of observing such t-value under H0 is about 0.38.

d. Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 :β1 =0?

```
fitted_regression_1 = lm (y ~ x1)
summary(fitted_regression_1)
```

```
##
## 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
```

After removing the predictor X2, we generate $\hat{\beta}_1 = 1.97$ which is much more close to the value of true value of $beta_1 = 2$ in the data set. The p-value of $\hat{\beta}_1$ is much more smaller than the threshold of 0.05, and we have strong evidence to reject the null hypothesis because the probability of obersving such t-value under null hypothesis is 2.66e-06.

e. Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis H0 :β1 =0?

```
fitted_regression_2 = lm (y ~ x2)
summary(fitted_regression_2)
```

```
##
## 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
```

After removing the predictor X1, we generate $\hat{\beta}_1 = 2.89,$ and the p-value of $\hat{\beta}_1$ is much more smaller than the threshold of 0.05, and we have strong evidence to reject the null hypothesis because the probability of observing such t-value under null hypothesis is 1.37e-05.

f. Do the results obtained in (c)–(e) contradict each other? Explain your answer. As we put both X_1 and X_2 into this regression model, both predictors appear not that statistically significant or even not significant is due to the high correlation(0.8) between these two predictor as they affect each other when they were both added to one linear regression model. However, when we split them into to linear regression line, they have strong relationship with the y response variable respectively, because the problem of multi-collinearity has been avoided ultimately. Therefore, the answer from section d is not contradict with section e.