Name: Chiayu Tu (Louis Tu)

Course: Stat 387 Assignment Two

Conceptual:

- **Carefully explain the differences between the KNN classifier and KNN regression methods.**

  The differences between KNN classifier and KNN regression is:

  KNN classifier attempts to predict the class to which the output variable belongs by computing the local probability. The KNN classifier shows Y as 0 or 1.

  KNN regression tried to predict the value of the output variable by using a local average. KNN regression method predicts the quantitative of Y and can also be continuous.

- **Suppose we have a data set with five predictors X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High school), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get β^0 = 50, , β^1 = 20, β^2 = 0.07, β^3 = 35, β^4 = 0.01, β^5 = −10.**
  - **Which answer is correct, and why?**
    - **For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.**
    - **For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.**
    - **For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.**
    - **For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.**

    X1 = GPA

    X2 = IQ

    X3 = Level (1: College, 0: High school)

    X4 = Interaction between GPA and IQ (X1, X2)

    X5 = Interaction between GPA and Level (X1, X3)

Using least square to fit the model, so we can get salary is:

Salary = B0 + B1 * 1 + B2 * 2 + B3 * 3 + B4 * 4 + B5 * 5

= 50 + 20 * 1 + 0.07 * 2 + 35 * 3 + 0.01 * 4 + (-10) * 5

= 50 + 20 + 0.14 + 105 + 0.04 – 50

= 125.18

Salary for high school

Salary = B0 + B1 * 1 + B2 * 2 + B3 * 0 + B4 * (X1, X2) + B5 * (X1, 0)

Salary for college

Salary = B0 + B1 * 1 + B2 * 2 + B3 * 1 + B4 * (X1, X2) + B5 * (X1, 1)

Salary (College) – Salary (High school)

= B0 + B1 * 1 + B2 * 2 + B3 * 1 + B4 * (X1, X2) + B5 * (X1, 1) – (B0 + B1 * 1 + B2 * 2 + B3 * 0 + B4 * (X1, X2) + B5 * (X1, 0))

= B3 * 1 + B5 * (X1, 1) – B5 * (X1, 0)

= 35 – 10 * (X1, 1) + 10 * (X1, 0)

= 35 – 10 * (X1 * 1) + 10 * (X1 * 0)

= 35 -10X1

- **Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.**

When salary of a college graduate with IQ of 110 and a GPA of 4.0
Salary = B0 + B1 * 4 + B2 * 110 + B3 * 1 + B4 * (4, 110) + B5 * (4, 1)
= 50 + 20 * 4 + 0.07 * 110 + 35 * 1 + 0.01 * (4 * 110) – 10 * (4 * 1)
= 137.1

Applied:

- **This question involves the use of simple linear regression on the Auto data set.**
  - **Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:**
    - **Is there a relationship between the predictor and the response?**

      There is a relationship between the predictor and the response because the p-value is 2.2e^-16.

    - **How strong is the relationship between the predictor and the response?**

      The R^2 value indicates that about 61% of the variation in the response variable (mpg) is due to the predictor variable (horsepower).

    - **Is the relationship between the predictor and the response positive or negative?**

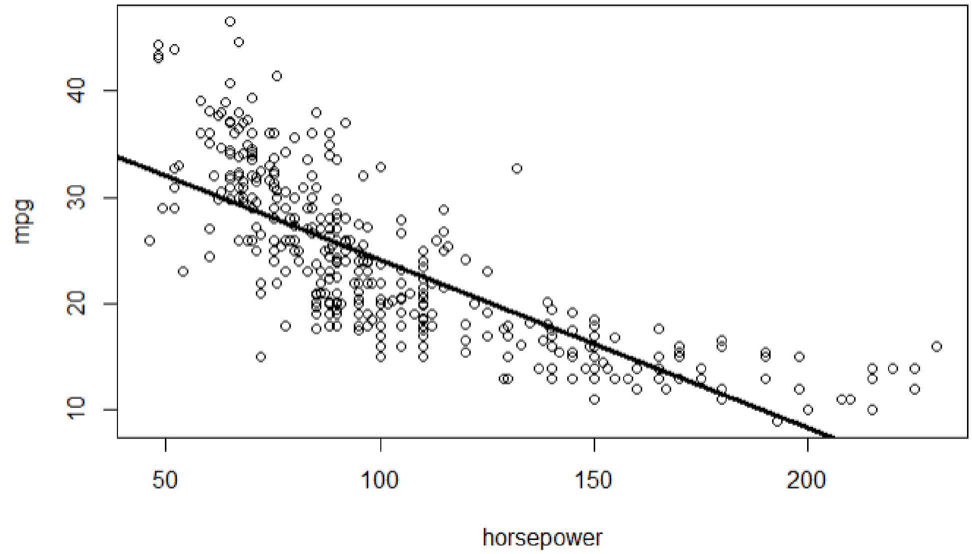
      Negative

    - **What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?**

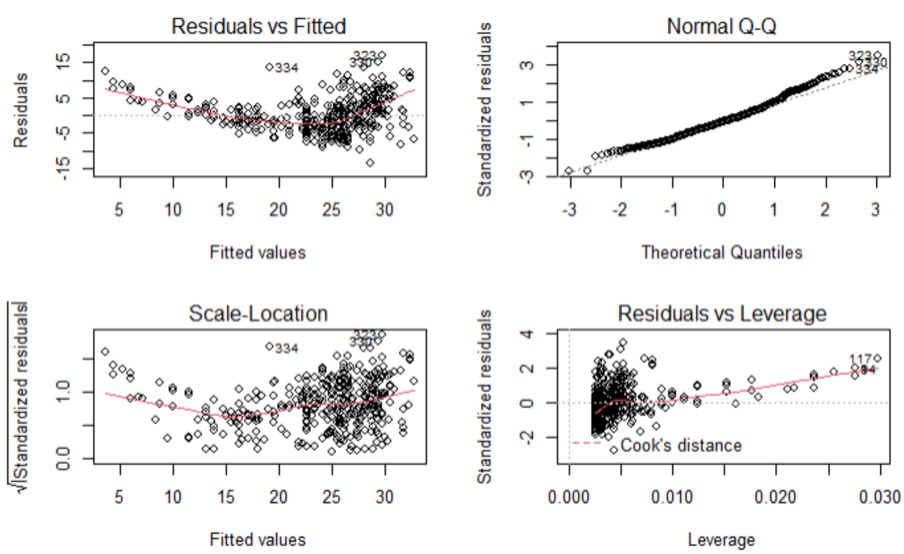
```
       fit     lwr      upr
1 24.46708 14.8094 34.12476
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5710  -3.2592  -0.3435   2.7630  16.9240 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 
1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049 
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- **Plot the response and the predictor. Use the abline() function to display the least squares regression line.**



- **Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.**



- **This question should be answered using the Carseats data set.**
  - **Fit a multiple regression model to predict Sales using Price, Urban, and US.**

```
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
```

  - **Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!**

    Price: From the summary above, we can observe that the relationship between price and sales given the low p-value. In the coefficient, the states of this two are negative relationship, which means that when price is increasing, sales is decreasing.

    UrbanYes: In the regression, the relationship between urban and sales shows that urban is not significant to sales. There is not evidence can prove that the location of the store can increase or decrease the sales.

    USYes: The relationship is shown positive between the store in the US and sales, which means that if the store is in the US, and the sales will increase by 1.200573.

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

    Sales = 13.043469 + (-0.054459) + (-0.021916) + 1.200573

- For which of the predictors can you reject the null hypothesis H0 : $\beta j = 0$?

  Price and USYes

- 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.

```
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
```

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

  The summary from (a) and (e), we can observe that model (e) is fitting the data slightly better than the model (a) though Residual standard error and R^2.

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

```r
x <- rnorm(100)
x
```

```
  [1]  0.28460754  0.61016853 -0.89052278  0.65226780
  [5]  0.81080327 -0.57522483  0.03010160  0.55890644
  [9]  0.66079002 -0.89468892  0.80387990  1.98433308
 [13] -0.39127710  0.12046731 -1.87701598 -0.22236471
 [17]  0.60735172 -0.70274812  0.54158897  0.05985390
 [21] -0.78028259  0.17042529  1.24328430  0.13461408
 [25] -0.29101854  1.12707420 -0.33661726 -0.71792477
 [29] -0.60343845 -0.40498740 -0.02641547  1.02501758
 [33]  1.31634849 -0.41029561  0.17911016 -1.47549322
 [37]  0.24746985 -0.10236873 -0.83084274 -1.01317601
 [41] -0.25521588 -0.57562494  0.90988439 -0.09803722
 [45] -0.73008791 -0.13228485 -0.69525441  0.08910638
 [49]  0.49609308  0.03875695  1.31653421 -1.27326545
 [53] -0.21871069  0.59274642 -0.78628525 -0.74130921
 [57] -0.40193851  1.17774004  1.08511993 -0.83155502
 [61] -1.46017714  1.95928998  1.08957921  0.60535709
 [65]  0.65570048  1.21939214 -0.23004217  0.98973541
 [69]  1.74553204  0.81361852 -0.36177684  1.20152920
 [73] -1.43516189  0.55286616 -0.52443335  0.16904050
 [77]  0.55162216  0.86467081 -2.04039634  0.65317543
 [81]  0.25996396  1.47944857  0.86736826  0.22112833
 [85]  0.04221492  1.32303095 -0.11689327 -0.76017465
 [89] -0.03503977  0.51490421  1.04895243 -0.78341441
 [93] -0.37004742  1.42912168 -0.78143408  0.10097358
 [97] -0.81154333  0.47765897  0.25499644 -0.67334954
```

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

```{r b3}
eps <- rnorm(100, 0, sqrt(0.25))
eps
```

```
  [1]  0.017094338  0.545847417 -0.406587278 -0.257393134
  [5] -0.606471068  0.425180981  0.637603120 -0.244187860
  [9]  0.188108179 -0.370501290  0.352304186 -0.062067839
 [13] -0.118743956  0.567577578  0.440550255  0.188936165
 [17]  0.329408812 -0.777257469  0.748227509  0.053473152
 [21] -0.082432172 -0.405518704  0.508230471  0.929317627
 [25] -0.090925651 -1.107804988  0.146034187 -0.623596131
 [29]  0.242891498 -0.143208759 -0.190075205  0.226790766
 [33] -0.740159742 -1.374280023  0.159528674 -0.520313356
 [37]  0.042516876 -0.194926298  0.624089402 -0.022889862
 [41]  0.567274391  0.781325524 -0.552360301 -0.812209999
 [45]  0.055004866 -1.320766525 -0.392948424 -1.074726870
 [49] -0.284572344  0.206882464  0.470343502 -0.007262132
 [53] -0.852697437  0.377765894 -0.079992041 -0.356039079
 [57]  0.310175049  0.764534195 -0.318013426  0.203877228
 [61] -0.013762436 -0.353290344 -0.730769873 -0.564744827
 [65]  0.531216234  0.278686363 -0.201088906  0.350834874
 [69]  0.209763807  0.832148576  0.131617673 -0.009581873
 [73] -0.265411088 -0.306510278  0.691958881 -0.830824978
 [77]  0.542485466 -0.114108806  0.139243938 -0.328347386
 [81] -0.470142648  0.033232533  0.283781425 -0.271039568
 [85]  0.111808204 -0.167534535 -0.512168146 -0.289617166
 [89] -0.271207101  0.521079127  0.001580861  0.044676586
 [93]  0.517076770  0.109625542 -0.692816390 -0.011847683
 [97] -0.578540301 -0.213353897  0.446516296 -0.115785142
```

- **Using x and eps, generate a vector y according to the model**

  $Y = -1+0.5X + \epsilon$. (3.39)

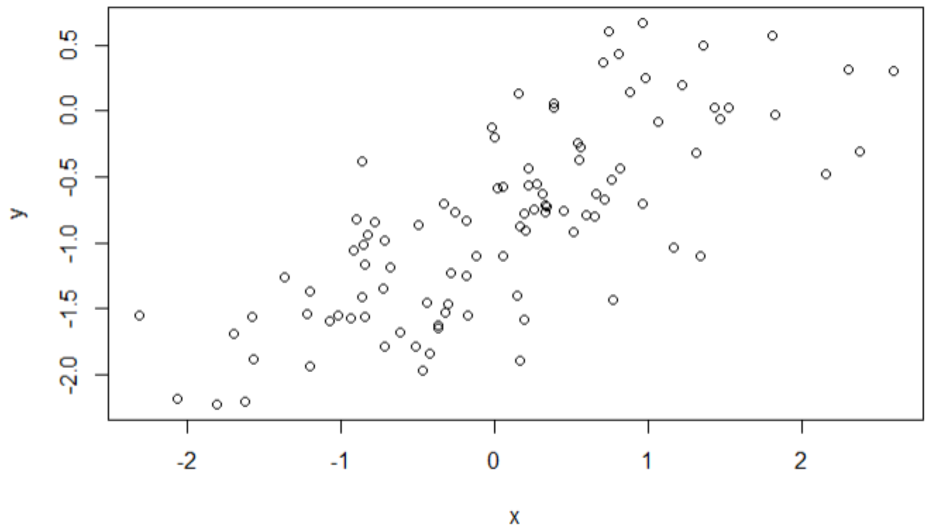
  What is the length of the vector y? What are the values of β and β1 in this linear model?

```{r c3}
y = -1 + 0.5 * x + eps
y
```

```
  [1] -2.19720539 -1.78608870 -0.47726247 -0.72403932
  [5] -0.57480360 -1.01760696 -0.76853940  0.02153307
  [9] -0.78577048 -0.55420640 -0.70339881 -1.61972972
 [13] -1.59477030 -1.26476597  0.49267943 -0.62953790
 [17]  0.24503628 -0.70377267 -0.37513727  0.67111497
 [21] -1.55321957 -0.82586579 -1.55651721  0.43132278
 [25] -0.58128550 -0.91850924  0.60289760 -0.19740589
 [29] -1.25278019 -0.86114129 -0.98353907 -0.06300583
 [33]  0.30025584 -2.18156581 -0.02784313 -1.96272048
 [37] -0.94141846 -2.22031354 -1.06141209 -0.51768534
 [41] -0.43338083 -1.67802567  0.14062838 -1.18562113
 [45] -1.09929866 -0.87022620 -0.67055521  0.06222937
 [49] -0.38218251 -0.75234877 -1.46434815 -0.30777938
 [53]  0.36298792 -0.84552893 -1.69042778 -0.90332133
 [57] -0.56803712 -1.41131025 -1.87941046 -1.34487375
 [61] -1.55314320 -1.55666375 -1.40236531 -0.27066801
 [65]  0.56860454 -1.52584882 -1.03546621 -1.55110904
 [69] -1.10195464 -1.93435387 -1.57840795 -0.71718912
 [73]  0.02366198 -1.09820655 -0.24579355 -0.43526854
 [77] -1.22682327  0.31339941 -0.62723127 -0.77534266
 [81] -0.12140167 -1.45227462 -0.31452072  0.13275827
 [85] -0.74056183  0.02060271 -1.89590371 -1.64555338
 [89] -1.36357315 -0.08091247 -0.82731967 -1.78138193
 [93] -1.16721883 -1.83871993  0.19336451 -1.42924985
 [97] -1.53445795 -1.56721149 -0.80095118 -0.76292142
```

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

```{r d3}
plot(x, y)
```



- **Fit a least squares linear model to predict y using x. Comment on the model obtained. How do β^0 and β^1 compare to β0 and β1?**

```{r e3}
model_xy <- lm(y ~ x)
summary(model_xy)
```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.10974 -0.27540 -0.02523  0.30561  1.07947 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.87539    0.04770  -18.35   <2e-16 ***
x            0.53486    0.04771   11.21   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4761 on 98 degrees of freedom
Multiple R-squared:  0.5619,	Adjusted R-squared:  0.5574 
F-statistic: 125.7 on 1 and 98 DF,  p-value: < 2.2e-16
```

R code:

```
## Question One

### a

library(ISLR)

library(MASS)

data(Auto)

head(Auto)

model_A <- lm(mpg ~ horsepower, data = Auto)

summary(model_A)

print("There is a relationship between the predictor and the response, because the p-value is 2.2e^-16.")

print("The R^(2) value indicates that aboout 61% of the variation in the response variable(mpg) is due to the predictor variable(horsepower)")

print("Negative")

predict(model_A, data.frame(horsepower = c(98)), interval = "prediction")

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


### b

attach(Auto)

plot(horsepower, mpg)

abline(model_A, lwd = 3)


### c

par(mfrow = c(2, 2))

plot(model_A)


## Question Two

### a
```

```
library(ISLR)
#head(Carseats)
#str(Carseats)
model_c = lm(Sales ~ Price + Urban + US, data = Carseats)
summary(model_c)

### e
model_c_1 <- lm(Sales ~ Price + US, data = Carseats)
summary(model_c_1)

## Question Three
### a
x <- rnorm(100)
x

### b
eps <- rnorm(100, 0, sqrt(0.25))
eps

### c
y = -1 + 0.5 * x + eps
y

### d
plot(x, y)

### e
model_xy <- lm(y ~ x)
summary(model_xy)
```