STA206 Fall 2022: Take Home Quiz

Instructions:

- In this quiz, you will be asked to perform some tasks in R
- You should submit a .html (preferred format) or .docx file.
- You should only include the output that is directly related to answering the questions. A flood of unprocessed raw output from R may result in penalties.

In *Quiz_data.Rdata* you will find a data set called *data* with three variables: *Y* and *X1*, *X2*. For the following, **you should use the original data and no standardization should be applied**.

- **(a). Load the data into the R workspace. How many observations are there in this data?**

```
#(Type your code in the space below)
my_data <- get(load("Quiz_data.RData"))
n = nrow(my_data)
print(n)
```

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

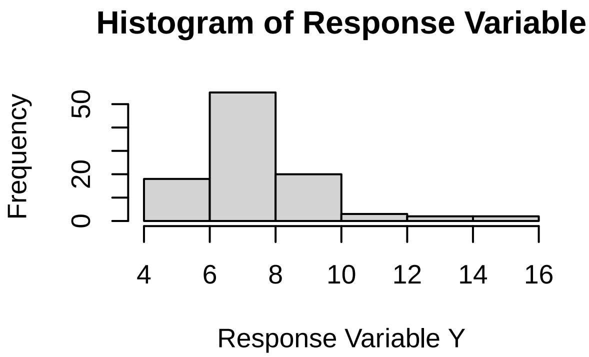
(Type your answer here):

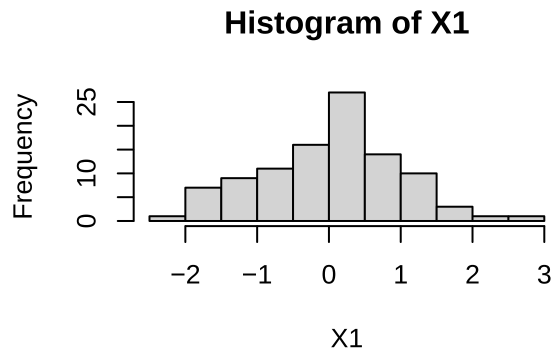
- **(b). What is the type of each variable? For each variable, draw one plot to depict its distribution. Arrange these plots into one multiple paneled graph.**

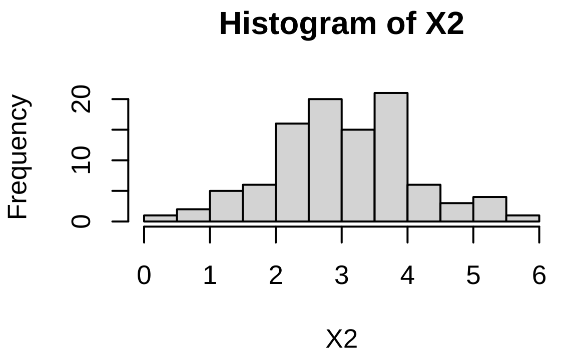
```
sapply(my_data,class)
```

```
##         Y        X1        X2
## "numeric" "numeric" "numeric"
```

```
par(mfrow = c(2, 2))
hist(my_data$Y, xlab='Response Variable Y', main='Histogram of Response Variable')
hist(my_data$X1, xlab='X1', main='Histogram of X1')
hist(my_data$X2, xlab='X2', main='Histogram of X2')
```



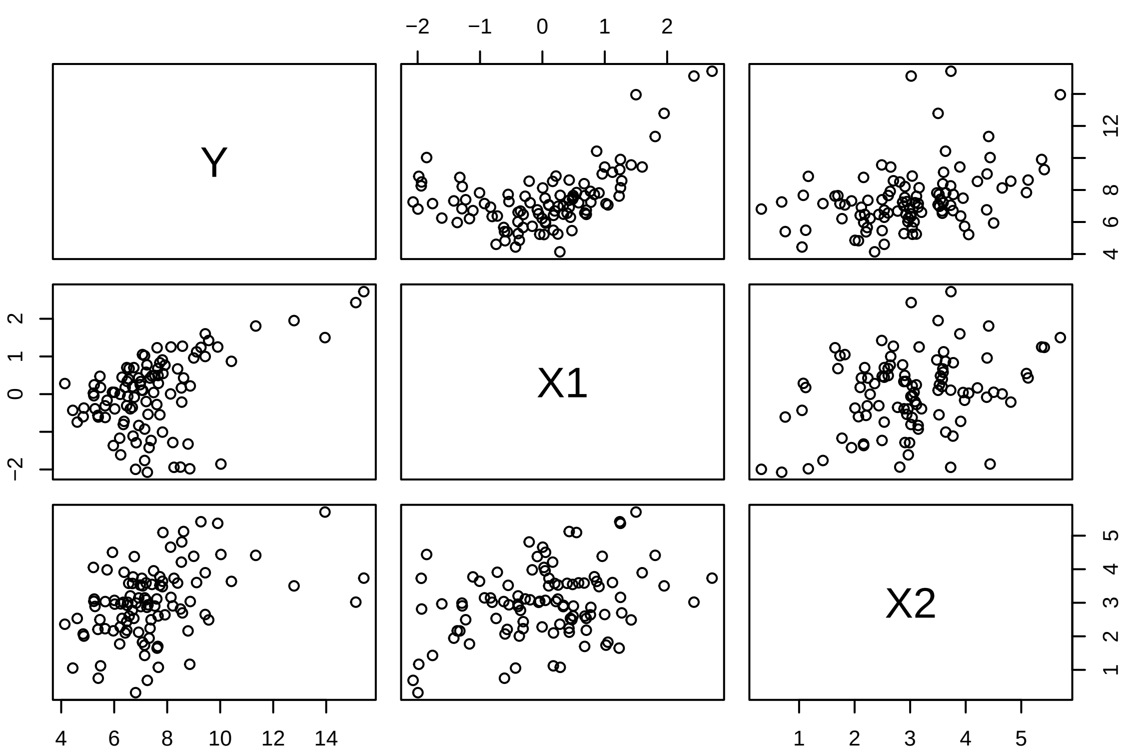




(Type your answer here):

- **(c). Draw the scatter plot matrix and obtain the correlation matrix for these three variables. Briefly describe how Y appears to be related to $X1$ and $X2$.**

```r
# (Type your code in the space below)
pairs(my_data)
```



It looks like Y is quadratically related to $X1$. The relationship is less clear with $X2$.

- **(d). Fit a first-order model with Y as the response variable and $X1$, $X2$ as the predictors (referred to as Model 1). How many regression coefficients are there in Model 1?**

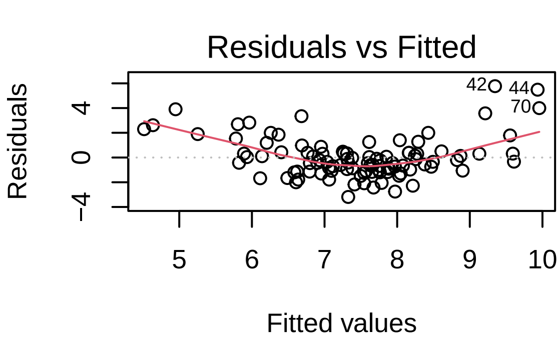
```
fit_1 = lm(Y ~ X1 + X2, data=my_data)
summary(fit_1)
```

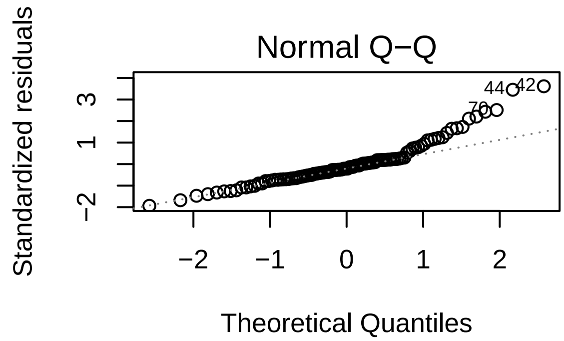
```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1897 -1.0656 -0.3424  0.3960  5.7706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9269     0.5246  11.298  < 2e-16 ***
## X1            0.7874     0.1764   4.464 2.17e-05 ***
## X2            0.4994     0.1662   3.005  0.00338 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.659 on 97 degrees of freedom
## Multiple R-squared:  0.3023, Adjusted R-squared:  0.288
## F-statistic: 21.02 on 2 and 97 DF,  p-value: 2.609e-08
```

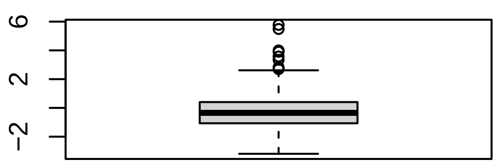
(Type your answer here): There are 3 regression coefficients.

- **(e). Conduct model diagnostics for Model 1 and comment on how well model assumptions hold.**

```
par(mfrow = c(2, 2))
plot(fit_1,which=1) ##residuals vs. fitted values
plot(fit_1,which=2) ##residuals Q-Q plot
boxplot(fit_1$residuals) ## residuals boxplot
```







(Type your answer here): We see that the residuals versus fitted line fails to be linear, and the Normal Q-Q plot has a rather large tail, which matches the outliers seen in the boxplot. So it's likely the linear assumptions will not hold well for this model.

- **(f). Fit a 2nd-order polynomial regression model with *Y* as the response variable and *X1*, *X2* as the predictors (referred to Model 2). Calculate the variance inflation factors for this model. Does there appear to be strong multicollinearity? Explain briefly.**

```r
fit_2 = lm(Y ~ X1 + X2 + I(X1^2) + X1:X2 + I(X2^2), data=my_data)
summary(fit_2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + X1:X2 + I(X2^2), data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20233 -0.60960 -0.07387  0.57877  2.31998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.25851    0.70035   7.508 3.39e-11 ***
## X1           0.93613    0.33911   2.761  0.00694 **
## X2           0.16454    0.46573   0.353  0.72467
## I(X1^2)      0.99757    0.07668  13.009  < 2e-16 ***
## I(X2^2)      0.06977    0.07475   0.933  0.35304
## X1:X2       -0.05632    0.11013  -0.511  0.61031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9371 on 94 degrees of freedom
## Multiple R-squared:  0.7844, Adjusted R-squared:  0.7729
## F-statistic: 68.38 on 5 and 94 DF,  p-value: < 2.2e-16
```

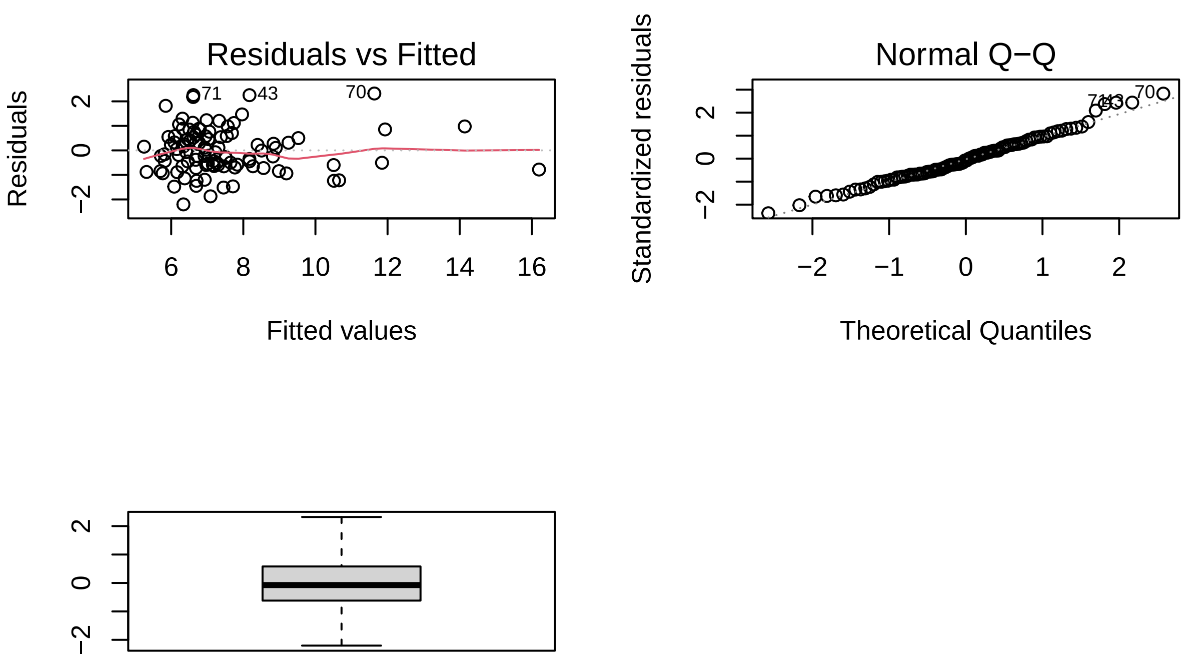
```
my_data['X1_squared'] <- (my_data$X1)^2
my_data['X1_X2'] <- (my_data$X1)*(my_data$X2)
my_data['X2_squared'] <- (my_data$X2)^2

correlation_matrix_including_y <- cor(my_data)
correlation_matrix <- correlation_matrix_including_y[2:6,2:6]

inverse_correlation_matrix <- solve(correlation_matrix)

VIF <- diag(inverse_correlation_matrix)
print(VIF)
```

```
##        X1        X2 X1_squared      X1_X2 X2_squared
## 12.942934 27.499045   1.251545 13.464178  27.772480
```

We see that the largest VIF factor is 27.772, which is much larger than 10, indicating that there is strong multicollinearity within our model.

- **(g). Conduct model diagnostics for Model 2. Do model assumptions appear to hold better under Model 2 compared to under Model 1? Explain briefly.**

```
par(mfrow = c(2, 2))
plot(fit_2,which=1) ##residuals vs. fitted values
plot(fit_2,which=2) ##residuals Q-Q plot
boxplot(fit_2$residuals) ## residuals boxplot
```



(Type your answer here):

Notice for the Normal Q-Q plot, we have a more controlled tail than that of model 1. This suggests that model 2 captures the nonlinear behavior better than the first model. We also see the residuals versus fitted lins splits the data as a line far better than model 1. So we could say that the model 2 assumptions hold better than those of model 1.

- **(h). Under Model 2, obtain the 99% confidence interval for the mean response when $X1 = X2 = 0$.**

```
#(Type your code in the space below)
alpha <- 1 - 0.99
hat_beta <- fit_2$coefficients
X_h = c(1, 0,0,0,0,0)
Y_h <- hat_beta %*% X_h
sqrt_MSE = sigma(fit_2)

ones <- rep(1, n)
X <- matrix(c(ones, my_data$X1, my_data$X2, my_data$X1_squared, my_data$X1_X2, my_data$X2_squared), nco
XtXinv <- solve( (t(X) %*% X) )
s_pred_h <- sqrt_MSE*sqrt(t(X_h) %*% XtXinv %*% X_h)

p <- 6
t_val <- qt(1 - alpha / 2, df = n - p)

confidence_intercal = c(Y_h - t_val*s_pred_h, Y_h + t_val*s_pred_h)
print(confidence_intercal)
```

```
## [1] 3.417190 7.099826
```

(Type your answer here):

$$CI_{99\%} = (3.417190, 7.099826)$$

- **(i). At the significance level 0.01, test whether or not all terms involving *X2* may be simultaneously dropped out of Model 2. State your conclusion.**

We develop the following hypothesis test:

- $H_0 : \beta_2 = \beta_4 = \beta_5 = 0$
- H_A : not all equal to zero
- Test Statistic: $F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$
- Null distribution of F^* is $F_{p-3,n-p}(\alpha) = F_{3,94}(\alpha)$

The rule becomes is we reject H_0 if the following comparison is true:

```
alpha = 0.01
SSE_F <- sum((fitted(fit_2) - my_data$Y)^2)
fit_3 = lm(Y ~ X1 + I(X1^2), data=my_data)
SSE_R <- sum((fitted(fit_3) - my_data$Y)^2)

F_stat <- ((SSE_R - SSE_F) / 3)/(SSE_F / (n - p))

F_crit <- qf(1 - alpha, 3, n-p, lower.tail = TRUE)
print(F_stat > F_crit)
```

```
## [1] TRUE
```

(Type your answer here):

Since the comparison is true, we conclude that we can that we can reasonably remove all terms involving X2.

- **(j) Find a model that has less regression coefficients AND a larger adjusted coefficient of multiple determination compared to Model 2. Briefly explain how you reach this model.**

```r
fit_4 = lm(Y ~ X1 + I(X1^2) + X2, data=my_data)
summary(fit_4)
```

```
##
## Call:
## lm(formula = Y ~ X1 + I(X1^2) + X2, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22047 -0.56627 -0.08829  0.53604  2.53136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.67108    0.30693  15.219  < 2e-16 ***
## X1           0.76504    0.09905   7.724 1.09e-11 ***
## I(X1^2)      0.99374    0.06830  14.551  < 2e-16 ***
## X2           0.59043    0.09352   6.313 8.46e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9316 on 96 degrees of freedom
## Multiple R-squared:  0.7824, Adjusted R-squared:  0.7756
## F-statistic:   115 on 3 and 96 DF,  p-value: < 2.2e-16
```

Here, the adjusted R squared is 0.7756 compared to 0.7729 with only 4 coefficients instead of 6. The previous question suggests that the information provided by all of the X2 terms is negligible, but keeping one can be valuable as seen by the higher coefficient.