STA206 Fall 2022: Take Home Quiz (Solution)

Instruction:

- In this quiz, you will be asked to perform some tasks in R
- You should submit a .html (preferred format) or .docx file

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?**

```r
#(Type your code in the space below)
load("Quiz_data.Rdata")
cat("There are", dim(data)[1], "observations in the data.")
```

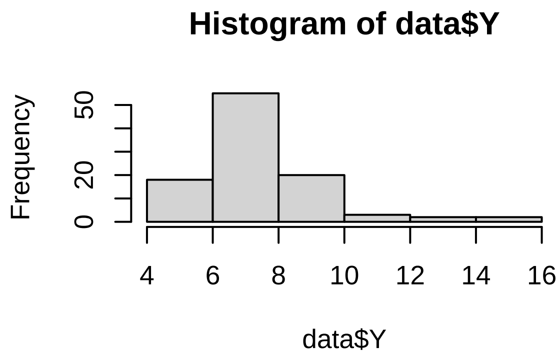
```
## There are 100 observations in the data.
```

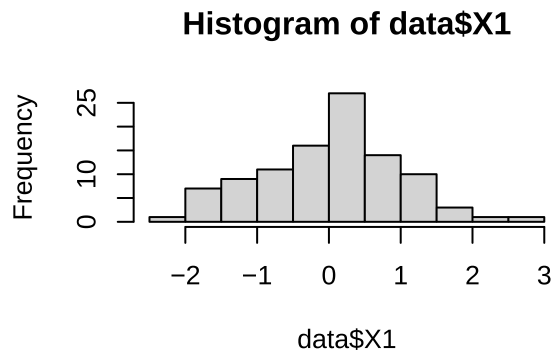
- **(b). What is the type of each variable? For each variable, draw one plot to depict its distribution. Arrange these plots into one overall graph.**

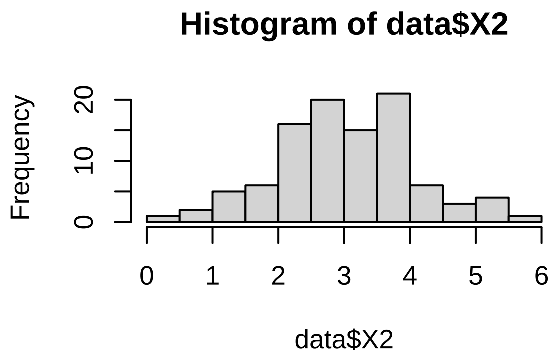
```r
#(Type your code in the space below)
sapply(data, class)
```

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

```r
par(mfrow=c(2,2))
hist(data$Y)
hist(data$X1)
hist(data$X2)
par(mfrow=c(1,1))
```



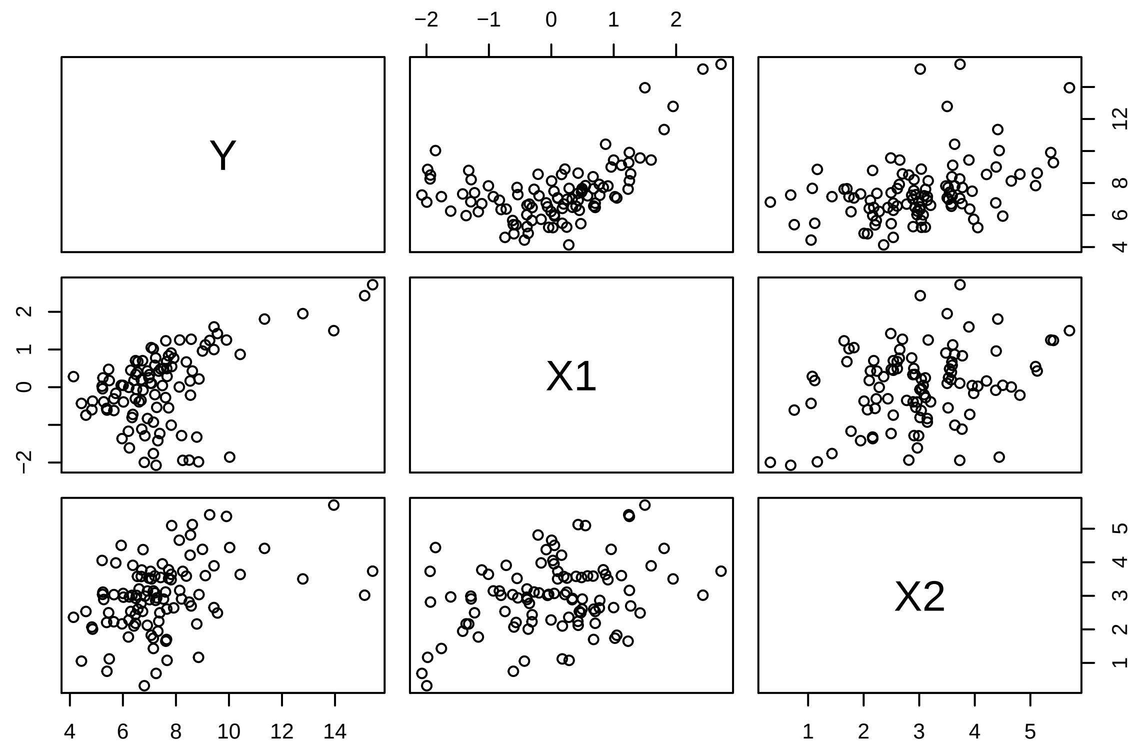




(Type your answer here): All three variables are of *numeric* type.

- **(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*.**

```
# (Type your code in the space below)
plot(data) ## alternatively: pairs(data)
```




```r
cor(data)
```

```
##            Y        X1        X2
## Y  1.0000000 0.4872571 0.3987890
## X1 0.4872571 1.0000000 0.3236289
## X2 0.3987890 0.3236289 1.0000000
```

Y is positively related to X_1 and X_2. The relationship between Y and X_2 appears to be linear, but there appears to be a nonlinear relationship between Y and X_1.

(Type your answer here):

- **(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?**

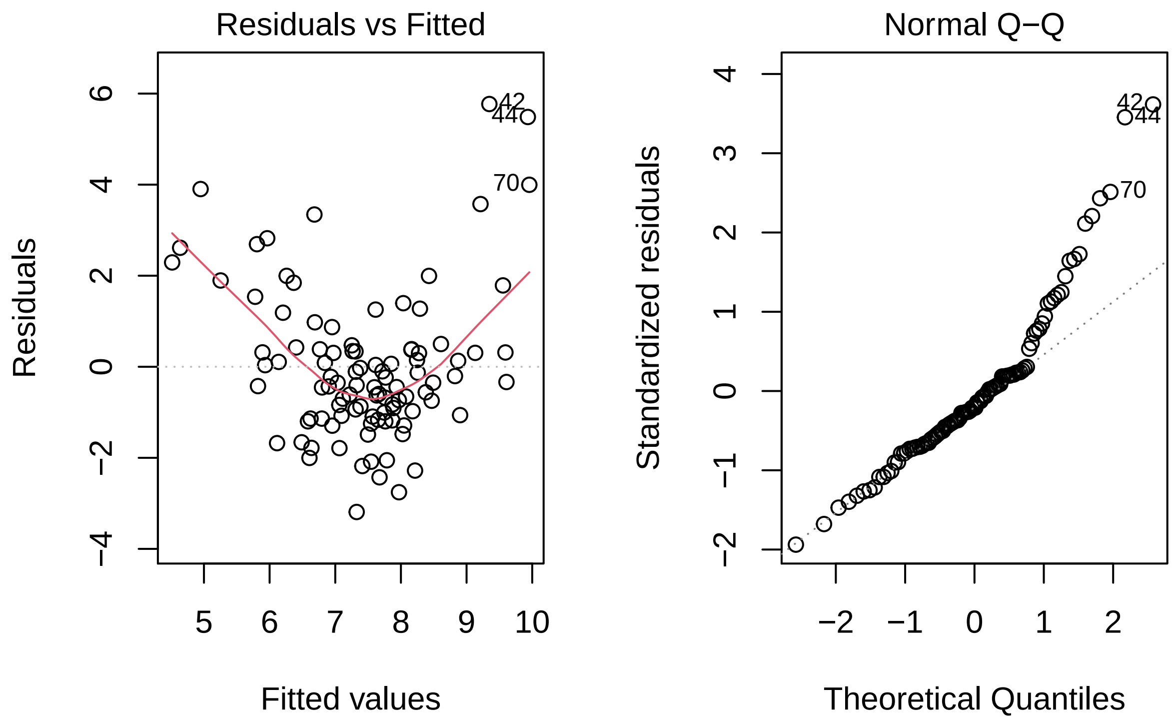
```r
# (Type your code in the space below)
fit1=lm(Y~X1+X2, data=data) ## alternatively:  lm(Y~., data=data)
```

(Type your answer here):

There are 3 regression coefficients (including the intercept) in Model 1.

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

```r
# (Type your code in the space below)
par(mfrow=c(1,2))
plot(fit1, which=1:2)
```




```r
par(mfrow=c(1,1))
```

(Type your answer here):

There is a nonlinear trend in the residuals vs. fitted values plot, which indicates the linearity assumption does not hold well. Moreover, the residuals Q-Q plot deviates from a straight line pattern. This might indicate that the normal error assumption is violated, but it could also be due to nonlinearity.

- **(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 appears to be strong multicollinearity? Explain briefly.**

```r
# (Type your code in the space below)
fit2=lm(Y~X1+I(X1^2)+X2+I(X2^2)+X1:X2,data=data)
diag(solve(cor(model.matrix(fit2)[,-1]))) ## alternatively, set up design matrix manually
```

```
##        X1   I(X1^2)        X2   I(X2^2)     X1:X2
## 12.942934  1.251545 27.499045 27.772480 13.464178
```

```r
## alternatively
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.5
```

```r
vif(fit2)
```

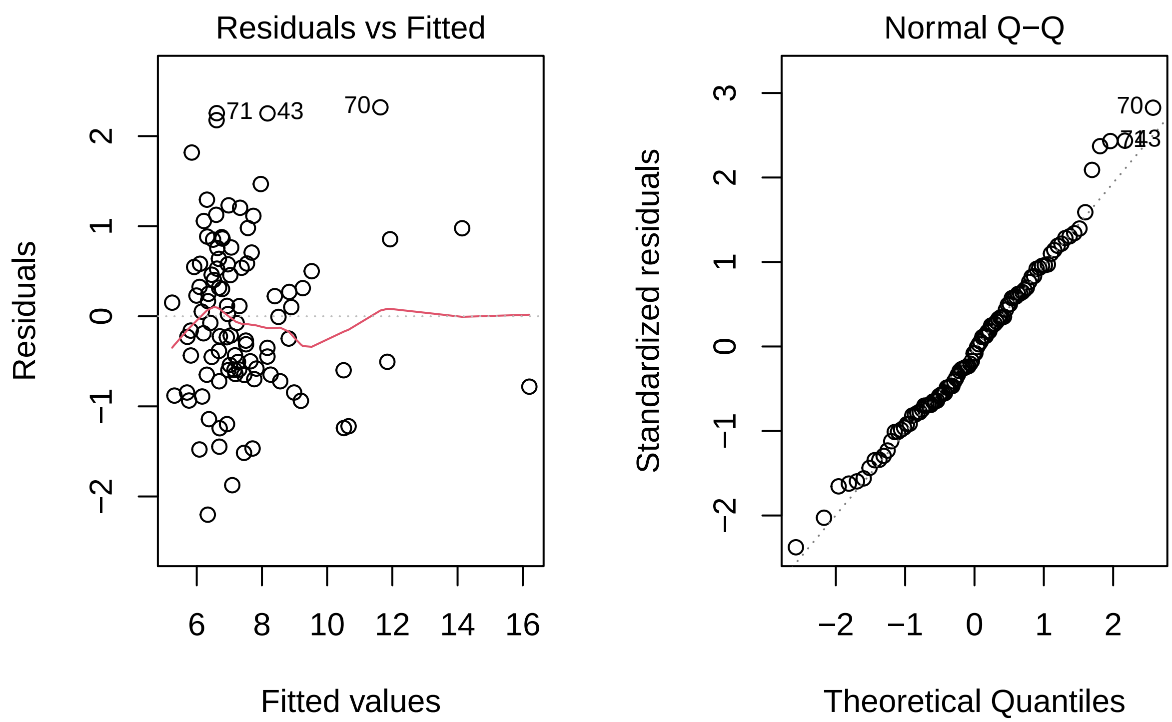
```
##        X1   I(X1^2)        X2   I(X2^2)     X1:X2
```

```
## 12.942934  1.251545 27.499045 27.772480 13.464178
```

(Type your answer here): The maximum VIF is 27 and greater than 10, which indicates fairly high multicollinearity among Model 2 X variables.

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

```r
# (Type your code in the space below)
par(mfrow=c(1,2))
plot(fit2, which=1:2)
```




```r
par(mfrow=c(1,1))
```

(Type your answer here):

Compared to residuals plots under Model 1, model assumptions appear to hold better under Model 2 as there is no obvious nonlinear trend in the residuals vs. fitted values plot and the residuals Q-Q plot follows a straight line pattern quite well.

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

```r
#(Type your code in the space below)
x.new=data.frame(X1=0,X2=0)
predict.lm(fit2, x.new, level=0.99, interval="confidence")
```

```
##        fit     lwr      upr
## 1 5.258508 3.41719 7.099826
```

(Type your answer here): The confidence interval is $[3.41719, 7.099826]$.

- **(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.**

```
#(Type your code in the space below)

fit3 = lm(Y~ X1 + I(X1^2), data = data)
anova(fit3, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + I(X1^2)
## Model 2: Y ~ X1 + I(X1^2) + X2 + I(X2^2) + X1:X2
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     97 117.899
## 2     94  82.541  3    35.358 13.422 2.307e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## alternatively
aov2=anova(fit2)
SSR.2.1=sum((aov2$'Sum Sq')[3:5])
Fstar=(SSR.2.1/3)/aov2$`Mean Sq`[6]
Fstar
```

```
## [1] 13.42233
```

```
1-pf(Fstar, 3,94)
```

```
## [1] 2.307253e-07
```

(Type your answer here): The p-value is $2.307e-07$ and is smaller than 0.01, so we can **not** simultaneously drop all terms involving *X2* out of Model 2.

- **(j) Find a model that has less regression coefficients and at the same time a larger adjusted coefficient of multiple determination compared to Model 2. Briefly state how you reach this model.**

```
#(Type your code in the space below)
fit4 = lm(Y~.^2 + I(X1^2), data = data) ## Adjusted R-squared: 0.7732
summary(fit4)
```

```
##
## Call:
## lm(formula = Y ~ .^2 + I(X1^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21958 -0.56348 -0.08906  0.53519  2.50797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.671860   0.308676  15.135  < 2e-16 ***
## X1          0.744665   0.269843   2.760  0.00694 **
## X2          0.590256   0.094036   6.277 1.03e-08 ***
## I(X1^2)     0.991038   0.076309  12.987  < 2e-16 ***
## X1:X2       0.007041   0.086669   0.081  0.93542
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9364 on 95 degrees of freedom
## Multiple R-squared:  0.7824, Adjusted R-squared:  0.7732
## F-statistic: 85.38 on 4 and 95 DF,  p-value: < 2.2e-16
```

```r
summary(fit2) ## Adjusted R-squared:   0.7729
```

```
##
## Call:
## lm(formula = Y ~ X1 + I(X1^2) + X2 + I(X2^2) + X1:X2, data = 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 **
## I(X1^2)      0.99757    0.07668  13.009  < 2e-16 ***
## X2           0.16454    0.46573   0.353  0.72467
## 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
```

```r
##alternatively
fit4=lm(Y~X1+I(X1^2)+X2, data=data)
summary(fit4) ##Adjusted R-squared:   0.7756
```

```
##
## Call:
## lm(formula = Y ~ X1 + I(X1^2) + X2, data = 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
```

(Type your answer here): From the scatter plot, there is no evidence that Y and X_2 have any non-linear relationship, so we may drop X_2^2 from Model 2. As a result, the adjusted R-squared is increased from 0.7729 to 0.7732. Another option is to drop both $X_2^2, X_1 * X_2$ from Model 2. The R_a^2 would be increased to 0.7756.