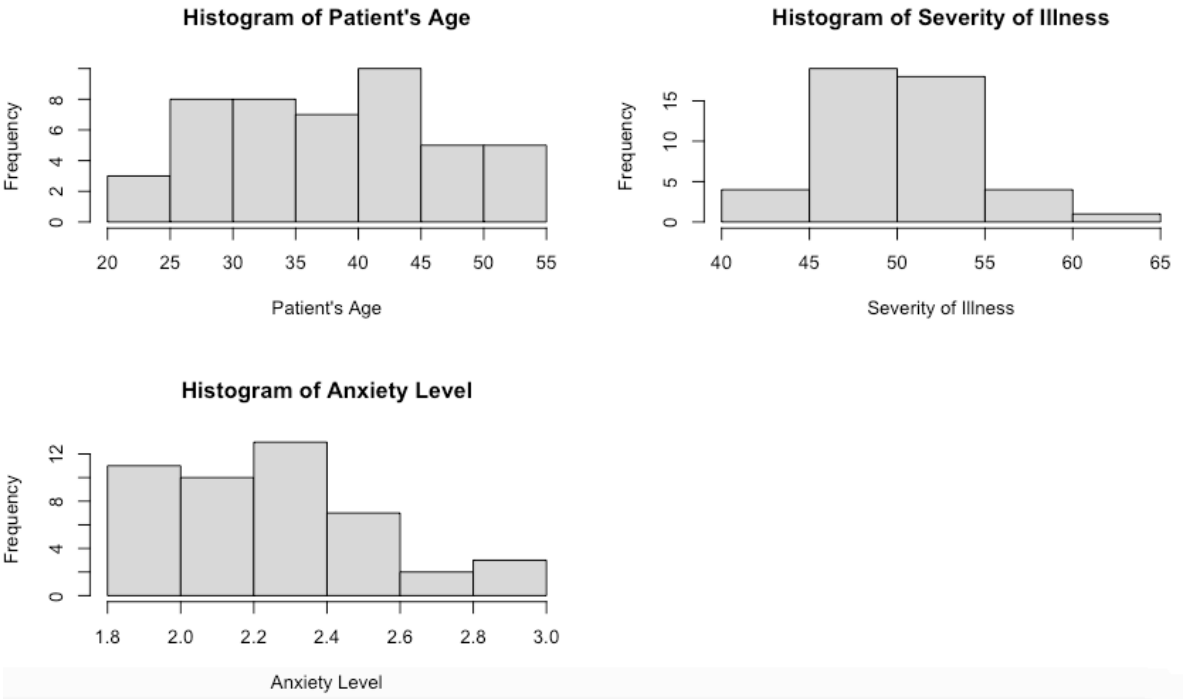
STP530 HW5 Solution

6.15

a.

Use the following R code to read in the dataset and create the histogram for each predictor variable:

```
setwd("~/Documents/ASU/STP530-YiZheng/HW-HaozhenXu/HW5/")

# Read data from txt file into R and assign data to an object called "HW5.data"

HW5.data <- read.table("CH06PR15.txt")

colnames(HW5.data) <- c("Y", "X1", "X2", "X3")

head(HW5.data)

# Create a histogram for each predictor variable

par(mfrow=c(2, 2)) # Create a graphing device that holds 2 rows and 2 columns
                   # of subfigures

hist(HW5.data$X1, xlab="Patient's Age", main="Histogram of Patient's Age")

hist(HW5.data$X2, xlab="Severity of Illness", main="Histogram of Severity of Illness")

hist(HW5.data$X3, xlab="Anxiety Level", main="Histogram of Anxiety Level")

dev.off() # Close the graphing device so that the next plot appears in a new figure
```

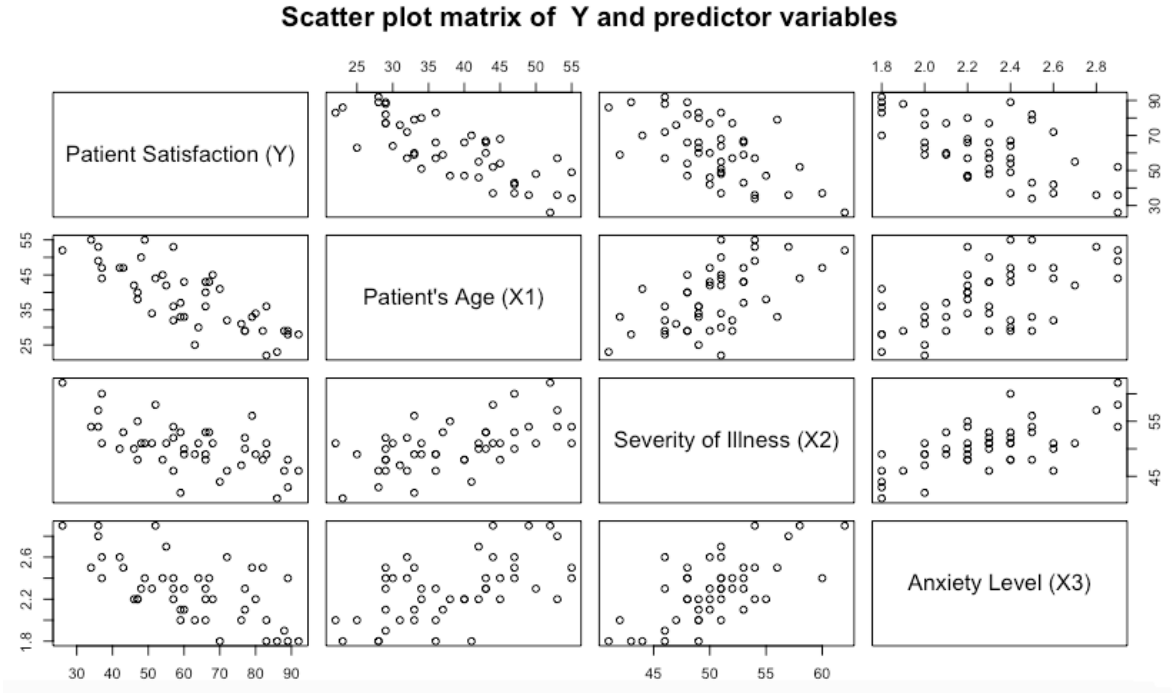


From these three histograms, we can see that all three predictors appear uni-modal. The distributions of X_1 (patient's age) and X_2 (severity of illness) are mostly symmetrical, while X_3 (anxiety level) has a right-skewed distribution. No obvious outlier is seen in the histograms. These distributions seem adequate to enter a regression analysis.

b.

Use the following code to obtain the scatter plot matrix and the correlation matrix:

```
# Scatter plot matrix

pairs(HW5.data, main="Scatter plot matrix of Y and predictor variables",
      labels = c("Patient Satisfaction (Y)",
                 "Patient's Age (X1)",
                 "Severity of Illness (X2)",
                 "Anxiety Level (X3)")
      )

# Correlation matrix

cor(HW5.data)
```




```
> cor(HW4.data[,1:4])
           Y         X1         X2         X3
Y   1.0000000 -0.7867555 -0.6029417 -0.6445910
X1 -0.7867555  1.0000000  0.5679505  0.5696775
X2 -0.6029417  0.5679505  1.0000000  0.6705287
X3 -0.6445910  0.5696775  0.6705287  1.0000000
```

The first row (and equivalently the first column) reveals the bivariate association between Y and each predictor variable. It appears that there are linear relationships between Y and each predictor variable, and the relationships are all negative, with a moderate strength.

Rows 2-4 with Columns 2-4 in the scatter plot matrix as well as the numbered correlation matrix depict the relationship among the predictors themselves. It appears there are moderate, positive linear relationships within every pair of the predictor variables. This is called a moderate multicollinearity problem, which will be discussed in the later weeks.

c.

Use the following code to fit regression model:

```
# Fit regression model

my.mod <- lm(Y ~ X1 + X2 + X3, data=HW5.data)

summary(my.mod)
```

```
> summary(my.mod)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = HW4.data)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3524  -6.4230   0.5196   8.3715  17.1601

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
X1           -1.1416     0.2148  -5.315 3.81e-06 ***
X2           -0.4420     0.4920  -0.898   0.3741
X3          -13.4702     7.0997  -1.897   0.0647 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-squared:  0.6822,    Adjusted R-squared:  0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

From the R output, we can obtain the estimated regression function as

$$\hat{Y} = 158.4913 - 1.1416X_1 - 0.4420X_2 - 13.4702X_3$$

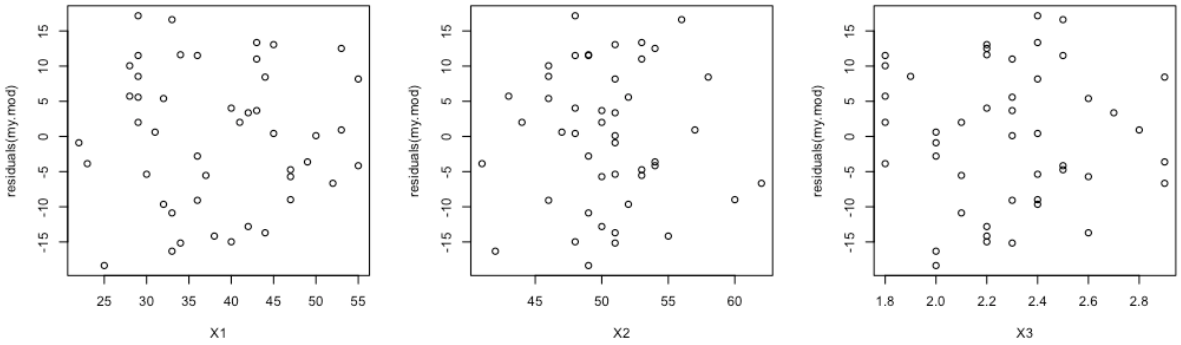
b_2 is the estimator of β_2, the slope of X_2. It means that if we keep X_1 (patient's age) and X_3 (anxiety level) unchanged, when X_2 (severity of illness) increases by one unit, the expected value of Y (patient satisfaction) decreases by 0.4420 unit. So this is the unique association between X_2 (severity of illness) and Y (patient satisfaction) after isolating out the influence of the other two predictors.

Added Question: Perform a full set of residual dianostics on the fitted model.

(1) Check whether the relationship between Y and each X is linear.

We plot the residuals against each X.

```
par(mfrow=c(1, 3))
plot(X1, residuals(my.mod))
plot(X2, residuals(my.mod))
plot(X3, residuals(my.mod))
```

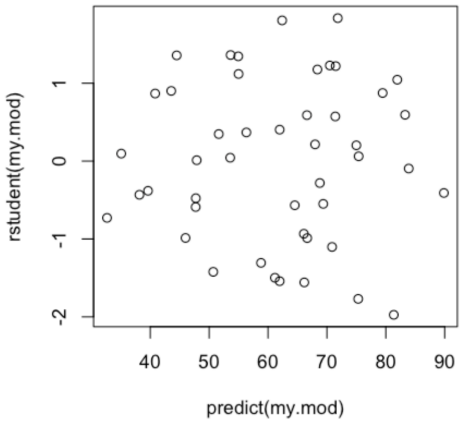


Impression: The residuals do not seem to relate to either X1, X2, or X3 in a systematic manner. So the first-order terms of the two predictors in model m seems sufficient.

(2) Check for outliers.

We plot the studentized residuals against Y-hat.

```
plot(predict(my.mod), rstudent(my.mod))
```

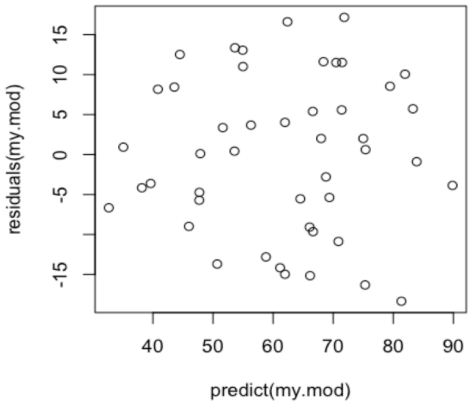


Impression: All studentized residuals are in a reasonable range given the approximate 1-2-3 rule. No excessive outliers are seen.

(3) Check for heteroskedasticity.

We plot the residuals against Y-hat.

```
plot(predict(my.mod), residuals(my.mod))
```

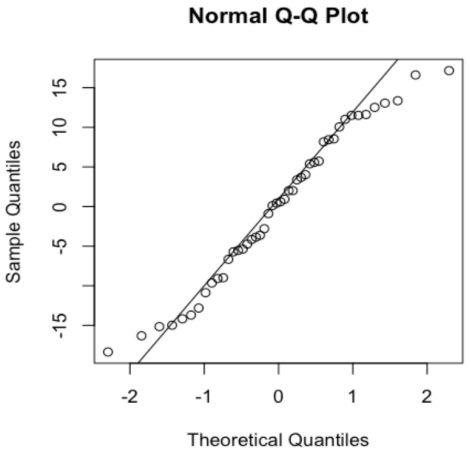


Impression: The vertical spread of the points are roughly constant across different X values. No concern of heteroskedasticity.

(4) Check whether the residuals are normally distributed.

We create the QQ-plot of the residuals.

```
qqnorm(residuals(my.mod))
qqline(residuals(my.mod))
```



Impression: The points are mostly aligned with the reference line. So the distribution of the residuals is approximately normal.

(5) Reflect whether the observations are independent from each other.

You would go to the problem description and the data collection method to find hints.

The problem states that the administrator randomly selected 46 patients, so we can reasonably assume the patients are independent from each other.

Summary:

This is a case where the data pass all basic diagnostic items easily. Now Step 1 of all the hypothesis tests in regression is formally completed. It is safe and valid to interpret the hypothesis test results.

6.16

a.

The five steps of the global F test of model utility:

- Assumptions:
$$\varepsilon \; i.i.d. \sim N(0, \sigma^2)$$
- Hypotheses:
$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad H_1 : not \; all \; \beta_k = 0 \; (k = 1, 2, 3)$$
- Test-statistic:

  There are two ways to find the test-statistic using R.

  (1) The first and easier way is to directly find it from the regression summary output shown on the previous page. You can find the "F-statistic", df1, df2, and p-value at the bottom of the output (last line).

  (2) The second way is to carry out the test-statistic formula using results provided by the ANVOA table:

```
> anova(my.mod)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 81.8026 2.059e-11 ***
X2         1  480.9   480.9  4.7539   0.03489 *
X3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F^* = \frac{MSR}{MSE} = \frac{SSR/df_{SSR}}{SSE/df_{SSE}} = \frac{(8275.4 + 480.9 + 364.2)/(1 + 1 + 1)}{101.2} = 30.05$$

Note that the numbers in the equation are found from the ANOVA table obtained above. The three rows of SS for predictors add up to be the global SSR; the three rows of df for predictors add up to be the df for SSR.

- P-value: Here the P-value is $P_{H_0}\{F^* > 30.05\} = 1.543485 \times 10^{-10}$. This p-value can be found from the R code below or directly from the regression summary output on page 4.

```
> 1-pf(30.05,3,42)
[1] 1.543485e-10
```

- Conclusion: Because the P-value $< \alpha = 0.1$, we can reject H_0 at a significance level of 0.1, which means there is at least one regression slope coefficient that is not equal to zero. In other words, at least one predictor is useful for predicting the outcome variable.

c.

We can obtain the coefficient of multiple determination from the regression summary output on page 4, which is 0.6822. It indicates that this regression model with the three predictor variables can interpret about 68.22% of the variation of Y.