STP 530: Applied Regression Analysis

Name : **Sai Swaroop Reddy Vennapusa**

Homework 5

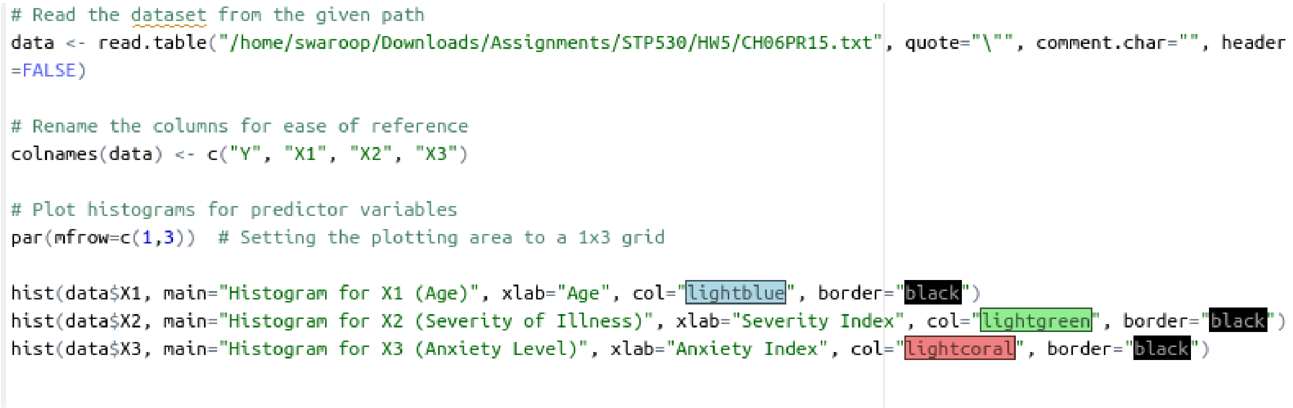
Instructor : **Yi Zheng**

Due Date : 26th Sep 2023, 10:30AM

Question 6.15: Patient satisfaction. A hospital administrator wished to study the l-elation between patient satisfaction (Y) and patient's age (X I, in years), severity of illness (X2, an index), and anxiety level (X 3 , an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of Y, X 2 , and X3 are, respectively, associated with more s,Hisfaction, increased severity of illness, and more anxiety.

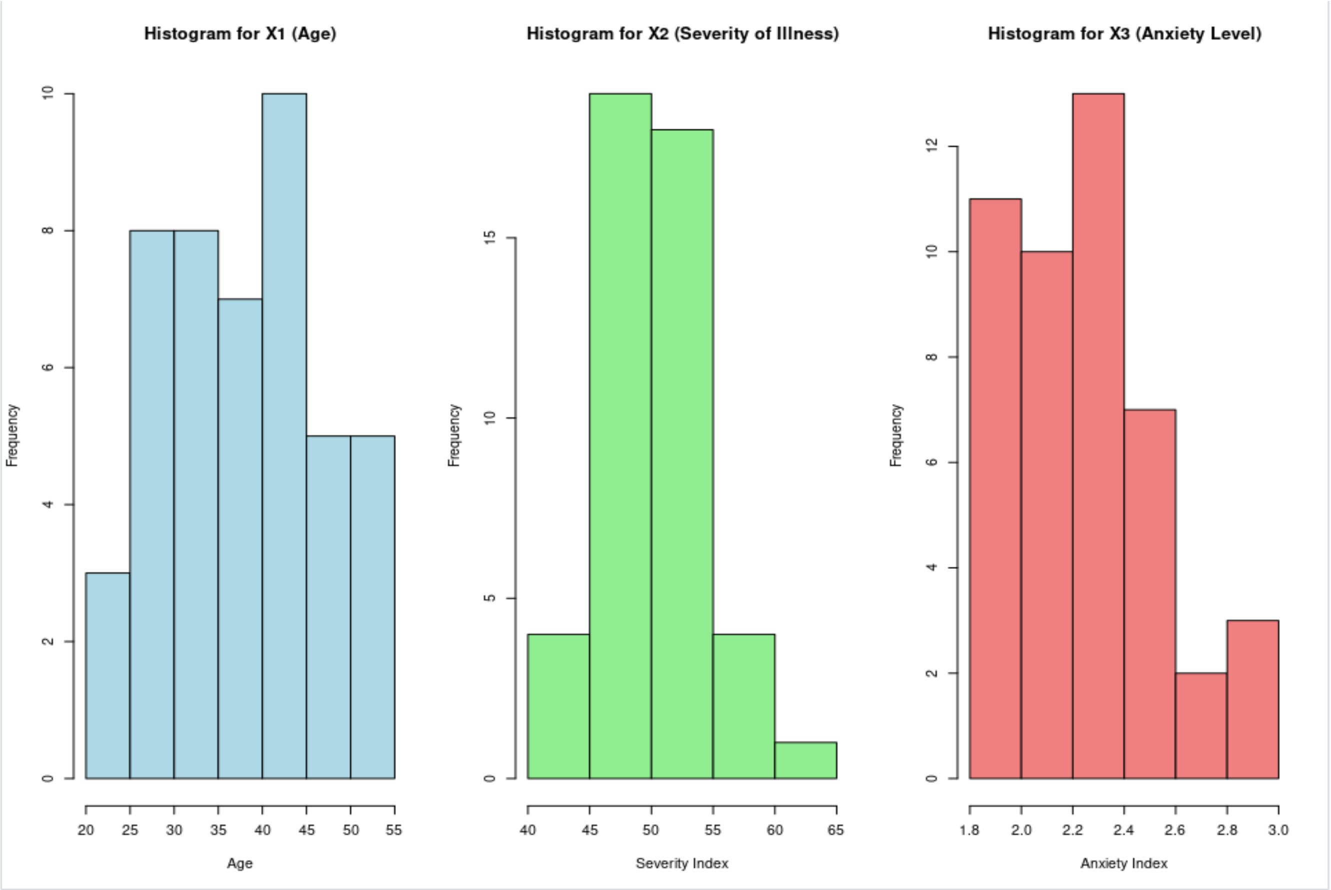
a) Prepare a stem-and-Ieaf plot for each of the predictor variables. Are any noteworthy features revealed by these plots? (Instructions: Instead of a stem-and-leaf plot, create a histogram for each predictor variable.)

Answer:
R Code:



```
# Read the dataset from the given path
data <- read.table("/home/swaroop/Downloads/Assignments/STP530/HW5/CH06PR15.txt", quote="\"", comment.char="", header
=FALSE)

# Rename the columns for ease of reference
colnames(data) <- c("Y", "X1", "X2", "X3")

# Plot histograms for predictor variables
par(mfrow=c(1,3))  # Setting the plotting area to a 1x3 grid

hist(data$X1, main="Histogram for X1 (Age)", xlab="Age", col="lightblue", border="black")
hist(data$X2, main="Histogram for X2 (Severity of Illness)", xlab="Severity Index", col="lightgreen", border="black")
hist(data$X3, main="Histogram for X3 (Anxiety Level)", xlab="Anxiety Index", col="lightcoral", border="black")
```

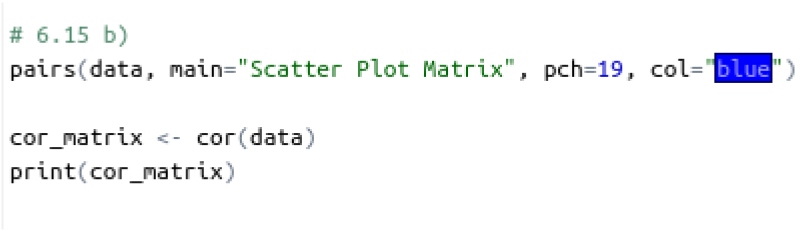
R Output:



b) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

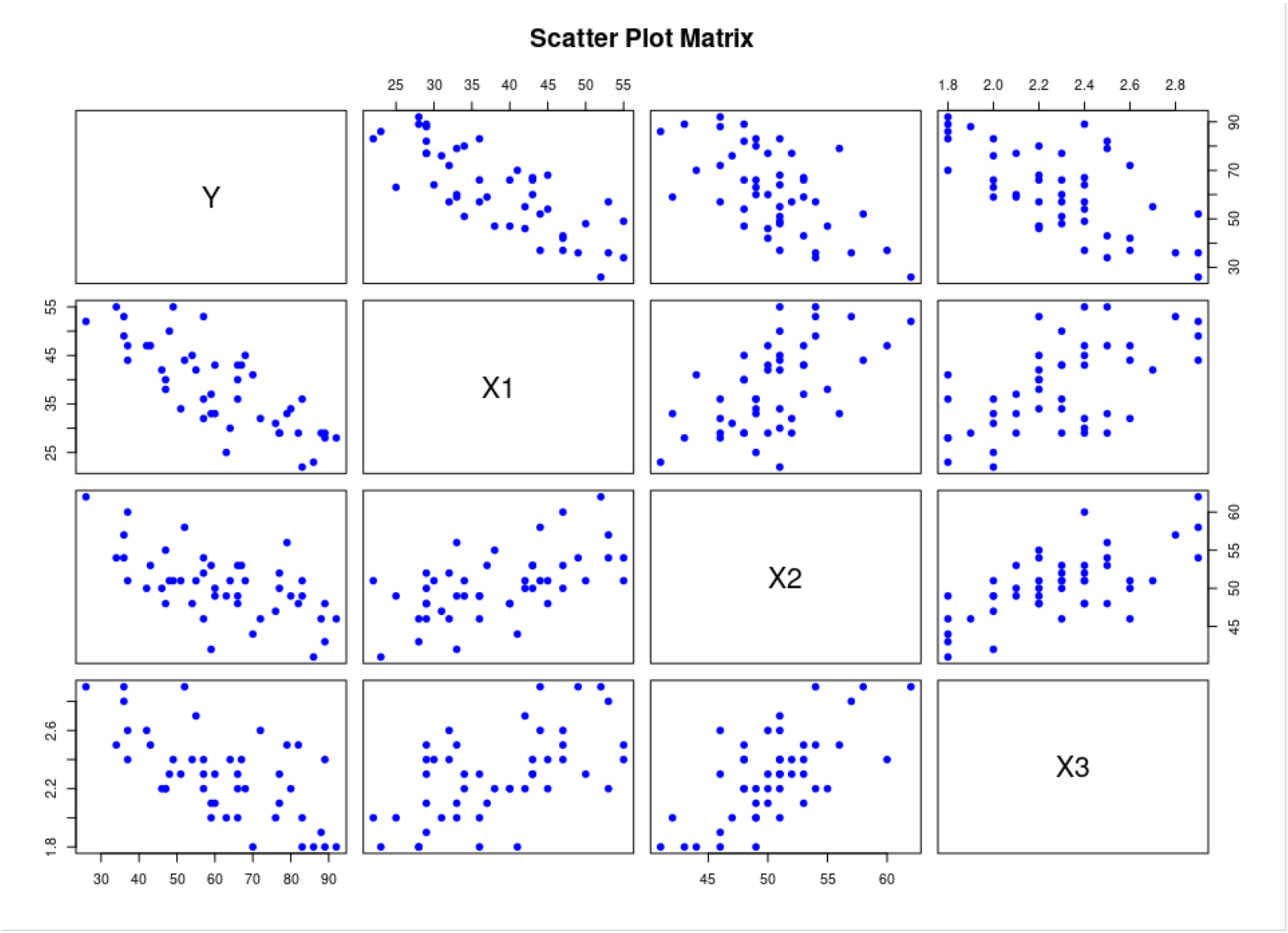
Answer:

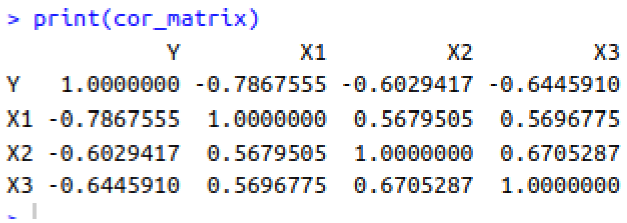
R code:



```
# 6.15 b)
pairs(data, main="Scatter Plot Matrix", pch=19, col="blue")

cor_matrix <- cor(data)
print(cor_matrix)
```

R Output:





```
> print(cor_matrix)
            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
```

Interpretation:

1. Y and X1: The correlation between Y and X1 is -0.7867555, indicating a strong negative linear relationship. As X1 increases, Y tends to decrease, and vice-versa.

2. Y and X2: The correlation between Y and X2 is -0.6029417, suggesting a moderate to strong negative linear relationship. An increase in X2 is associated with a decrease in Y.

3. Y and X3: The correlation coefficient between Y and X3 is -0.6445910, pointing to a moderate to strong negative linear relationship. As X3 goes up, Y tends to go down.

4. X1 and X2: The correlation between X1 and X2 is 0.5679505, indicating a moderate positive linear relationship. X1 tends to increase as X2 does.

5. X1 and X3: The correlation between X1 and X3 is 0.5696775, also signifying a moderate positive linear relationship.

6. X2 and X3: The correlation between X2 and X3 is 0.6705287, representing a strong positive relationship. X2 and X3 tend to move in the same direction.

Principal Findings:

1. Y (Patient Satisfaction) Decreases with Increase in X1 (Age): Older patients tend to report lower satisfaction.

2. Y Decreases with Increase in X2 (Severity of Illness) and X3 (Anxiety Level): Patients with higher severity of illness or higher anxiety levels report lower satisfaction.

3. Age (X1) Positively Correlates with Severity of Illness (X2) and Anxiety Level (X3): Older patients tend to have a higher severity of illness and anxiety level.

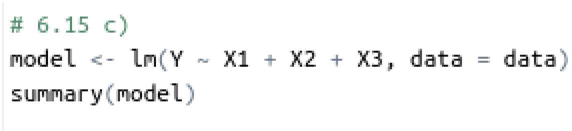
4. Severity of Illness (X2) and Anxiety Level (X3) are Positively Correlated: Patients with higher severity of illness tend to also have higher anxiety levels.

These findings provide insights into factors affecting patient satisfaction. For instance, patient satisfaction might be improved by focusing on reducing the anxiety levels of patients, especially among older patients or those with severe illnesses.

c) Fit regression model (6.5) for three predictor variables to the data and state the estimated regression function. How is b2 interpreted here?

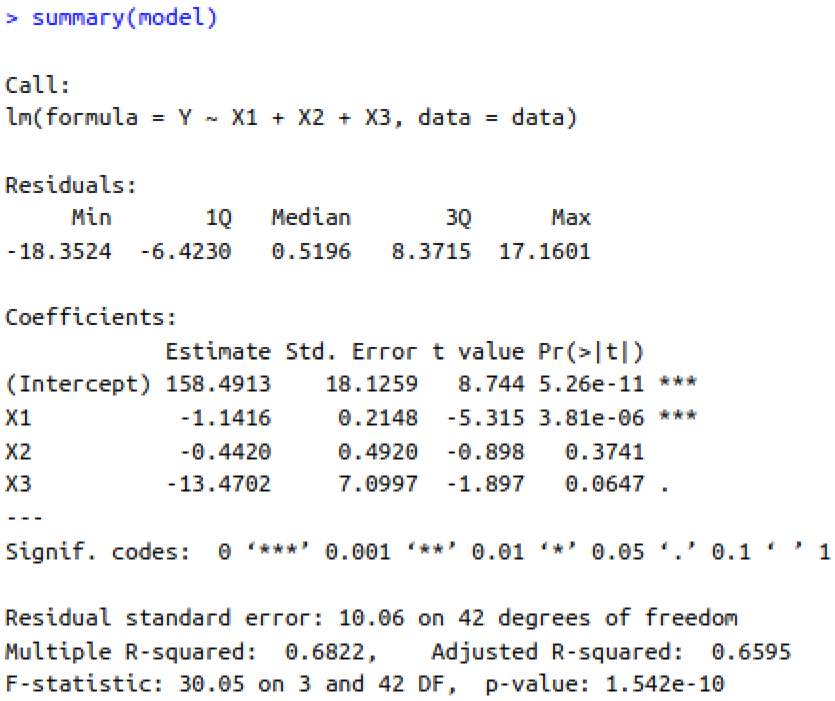
Answer:

R code:



```
# 6.15 c)
model <- lm(Y ~ X1 + X2 + X3, data = data)
summary(model)
```

R Output:



```
> summary(model)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = 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
```

The estimated regression function is given by:

$Y = 158.4913 – 1.1416X_1 – 0.4420X_2 – 13.4702X_3$

The coefficient b_2 (associated with X_2 , which represents severity of illness) is -0.4420.

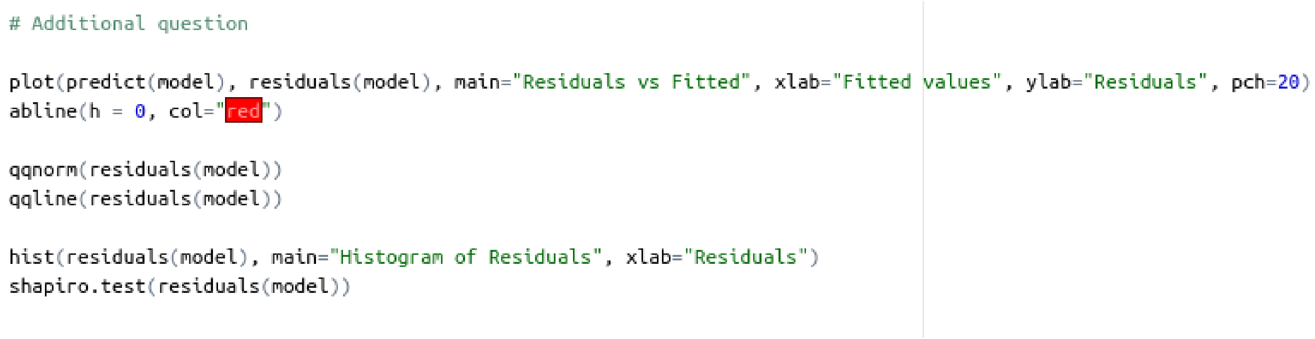
Interpretation of b_2 :
For every one-unit increase in the severity of illness (while holding the patient's age X_1 and anxiety level X_3 constant), the patient satisfaction (Y) is predicted to decrease by approximately 0.4420 units.

However, it's crucial to note that this coefficient is not statistically significant at the conventional 0.05 significance level, given its p-value of 0.3741. This means that, based on the data at hand, we do not have strong evidence to conclude that the severity of illness has a significant linear effect on patient satisfaction when accounting for age and anxiety level.

Additional Question: Follow examples in the diagnostics demo R code to conduct diagnostics and reflect on to what extent the sample data support that the assumptions of the normal error regression (NER) model (i.e.,ε_i iid $\sim N(0, \sigma^2)$) is reasonable.

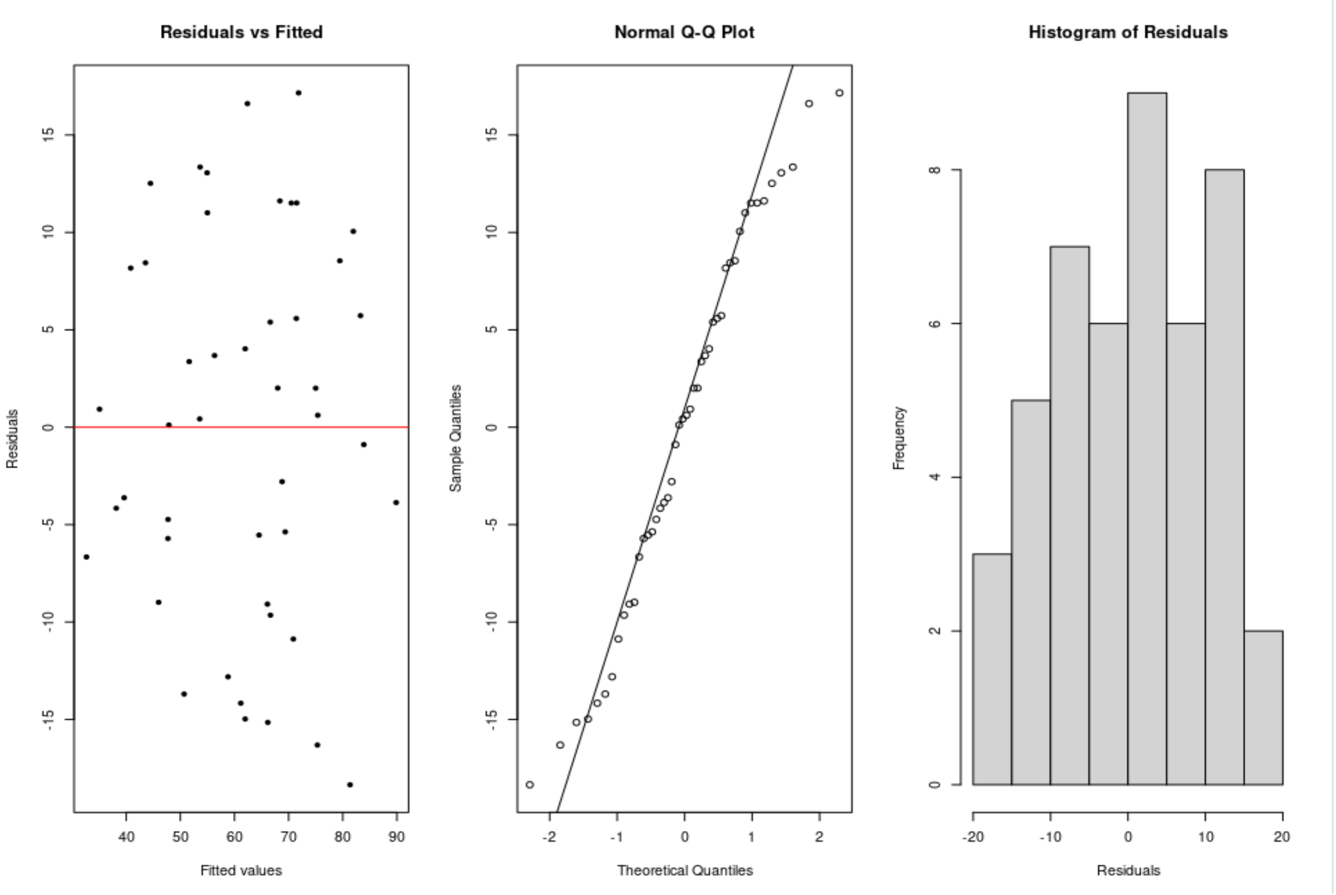
Answer:

R code:

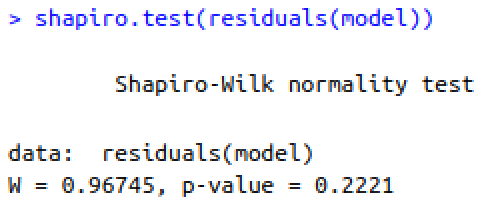


```
# Additional question

plot(predict(model), residuals(model), main="Residuals vs Fitted", xlab="Fitted values", ylab="Residuals", pch=20)
abline(h = 0, col="red")

qqnorm(residuals(model))
qqline(residuals(model))

hist(residuals(model), main="Histogram of Residuals", xlab="Residuals")
shapiro.test(residuals(model))
```

R Output:



Shapiro test:



```
> shapiro.test(residuals(model))

	Shapiro-Wilk normality test

data:  residuals(model)
W = 0.96745, p-value = 0.2221
```

Interpretation:

From the graphical methods:

1. Scatter Plot of Residuals vs. Fitted Values suggests that:
    - Linearity assumption is likely met, as the residuals are spread fairly randomly above and below the line.
    - Homoscedasticity assumption (constant variance of residuals) seems to be met as the spread remains fairly constant across the range of fitted values.

2. QQ-Plot:
    - Indicates that residuals are approximately normally distributed, as most of the points lie close to the line.
    - There are a few outliers, which may raise some concern about perfect normality.

From the statistical test:

- Shapiro-Wilk Normality Test:
    - The p-value is 0.2221, which is greater than 0.05. Thus, according to this test, the residuals don't deviate significantly from normality.

So, in summary:

- Both the graphical methods and the Shapiro-Wilk test suggest that the residuals are approximately normally distributed.
- The scatter plot suggests that the assumptions of linearity and homoscedasticity are met.

Hence, there's no contradiction between the graphical diagnostics and the statistical test. Both indicate that the assumptions for the linear regression model are reasonably met.

Question 6.16: Refer to Patient satisfaction Problem 6.15. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate.

a) Test whether there is a regression relation; use alpha = .10. State the alternatives, decision rule, and conclusion. What does your test imply about beta_l, beta_2, and beta_3? What is the P-value of the test?

Answer:

1. Assumptions:
For the regression model Y_i = beta_0 + beta_1X_{i1} + beta_2X_{i2} + beta_3X_{i3} + epsilon_i , the errors, epsilon , are assumed to be independent and identically distributed (i.i.d.) following a normal distribution with mean 0 and constant variance sigma^2 .

2. Hypotheses:
For the global test of a multiple regression model, the hypotheses are:
H_0: beta_1 = beta_2 = beta_3 = 0
H_1: At least one beta_j != 0 for j = 1, 2, 3

3. Test-statistic:
The observed value of the F-statistic, F_{obs}, is given by:
F_{obs} = MSR/MSE
From the regression summary, F_{obs} is 30.05.

4. P-value:
The p-value associated with the F-statistic is the probability of observing a value as extreme as F_{obs} under the null hypothesis. The p-value from summary is 1.542e-10.

5. Conclusion:
Given a significance level alpha = 0.10:

- If p-value < alpha, reject H_0.
- If p-value ≥ alpha, fail to reject H_0.

Since p-value = 1.542e-10 which is less than alpha = 0.10, the null hypothesis H_0 is rejected. This suggests that there is sufficient evidence to conclude that at least one of the predictor variables (X1, X2, or X3) is related to the response variable Y in the population.

c) Calculate the coefficient of multiple determination. What does it indicate here?

Answer:

The coefficient of multiple determination, commonly known as R^2 , represents the proportion of the variance in the dependent variable that's explained by the independent variables in a multiple regression model.

From the regression summary you provided:
R^2 = 0.6822

This means that approximately 68.22% of the variability in the dependent variable (Patient satisfaction, Y) is explained by the linear regression model which includes the three predictor variables (patient's age X1 , severity of illness X2 , and anxiety level X3).

In this context, the R^2 value of 0.6822 indicates that the three predictors combined explain a substantial portion (68.22%) of the variance in patient satisfaction. This suggests that the model fits the data fairly well, and these predictors are important in determining the patient satisfaction score. However, it's also worth noting that approximately 31.78% of the variability in Y is not explained by the model, which could be attributed to other unaccounted factors or inherent variability.