**STP 530: Applied Regression Analysis**

Name : **Sai Swaroop Reddy Vennapusa**

Homework **4**

Instructor : **Yi Zheng**

Due Date : $19^{th}$ Sep 2023, 10:30AM

Question 2.25: Refer to Airfreight breakage Problem 1.21.
(a.) Set up the ANOVA table. Which elements are additive?

Answer:

$p = 2$
$n = 10$

| Source | SS (Sum of Squares) | d.f. | MS (Mean Squares) |
|---|---|---|---|
| Regression | SSR = 160 | p−1 | MSR = 160 |
| Error | SSE = 17.6 | n-p | MSE = 2.2 |
| Total | SSTO = 177.6 | n-1 | MSTO = 19.733 |

R Code:

```
X <- c(1.0, 0.0, 2.0, 0.0, 3.0, 1.0, 0.0, 1.0, 2.0, 0.0)
Y <- c(16.0, 9.0, 17.0, 12.0, 22.0, 13.0, 8.0, 15.0, 19.0, 11.0)
model <- lm(Y ~ X)
summary(model)
Y_hat <- predict(model)
Y_bar <- mean(Y)
SSR <- sum((Y_hat - Y_bar)^2)
SSE <- sum((Y - Y_hat)^2)
SSTO <- sum((Y - Y_bar)^2)
p <- 2  # number of parameters (intercept + slope)
n <- length(Y)

MSR <- SSR / (p - 1)
MSE <- SSE / (n - p)
MSTO <- SSTO / (n - 1)

# Question 2.25(a)
anova(model)
```

R Output:

```
> anova(model)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1  160.0   160.0  72.727 2.749e-05 ***
Residuals  8   17.6     2.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b.) Conduct an F test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the a risk at .05. State the alternatives, decision rule, and conclusion.

Answer:

1. Assumptions:
   - The errors $\varepsilon$ are independent and identically distributed (i.i.d.) and follow a Normal distribution with mean 0 and variance sigma^2 .

2. Hypotheses:
   - H_0 : beta_1 = 0 (There is no linear association between the number of times a carton is transferred and the number of broken ampules.)
   - H_1 : beta_1 $\neq$ 0 (There is a linear association between the number of times a carton is transferred and the number of broken ampules.)

3. Test-statistic:
   - Given the ANOVA table, F_{obs} = MSR/MSE = 160/2.2 = 72.73

4. P-value:
   - Using the F_{obs} value of 72.73 and the degrees of freedom from the ANOVA table (df for regression = p-1 = 1 and df for error = n-p = 8), we can find the p-value.

R code :

```
# Question 2.25(b)
p_value <- 1 - pf(72.73, 1, 8)
```

R output:

```
> p_value
[1] 2.748294e-05
```

5. Conclusion:

   - The p-value 2.748294 times 10^{-5} is much less than the significance level alpha = 0.05 .
   - Therefore, we reject the null hypothesis H_0 .
   - This means there is significant evidence to suggest a linear association between the number of times a carton is transferred and the number of broken ampules.

(d.) Calculate R2 and r. What proportion of the variation in Y is accounted for by introducing -X into the regression model?

Answer:

R code:

```
# Question 2.25(d)
r <- cor(X, Y)
```

R output:

```
> r
[1] 0.949158
```

Given the data and ANOVA table details you've provided, we can calculate $R^2$ and $r$ (Pearson's correlation coefficient).

1. Calculation of $R^2$ :

$R^2 = SSR/SSTO$

From your ANOVA table:
$SSR = 160$
$SSTO = 177.6$

Plugging in the values, we get:
$R^2 = {160}/{177.6}$
$R^2 \approx 0.9011$

This means approximately 90.11% of the variation in $Y$ is explained by the linear regression model with $X$ as a predictor.

2. Calculation of $r$ (Pearson's correlation coefficient):

For simple linear regression:
$r = \sqrt{R^2}$
And the sign of $r$ will be the same as the sign of the slope $\beta_1$ (which we assume to be positive since the number of broken ampules would likely increase as the number of times a carton is transferred increases).

Given our calculated $R^2$ :
$r = \sqrt{0.9011}$
$r \approx 0.9493$

Therefore, the correlation coefficient r between X and Y is approximately 0.9493.

Interpretation:
About 90.11% of the variation in the number of broken ampules (Y) is accounted for by the number of times a carton is transferred (X) in the regression model. The high correlation coefficient of approximately 0.9493 further indicates a strong linear relationship between the two variables.

Question 3.25: Refer to the CDI data set in Appendix C.2 and Project 1.43. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?

Answer:

R code:

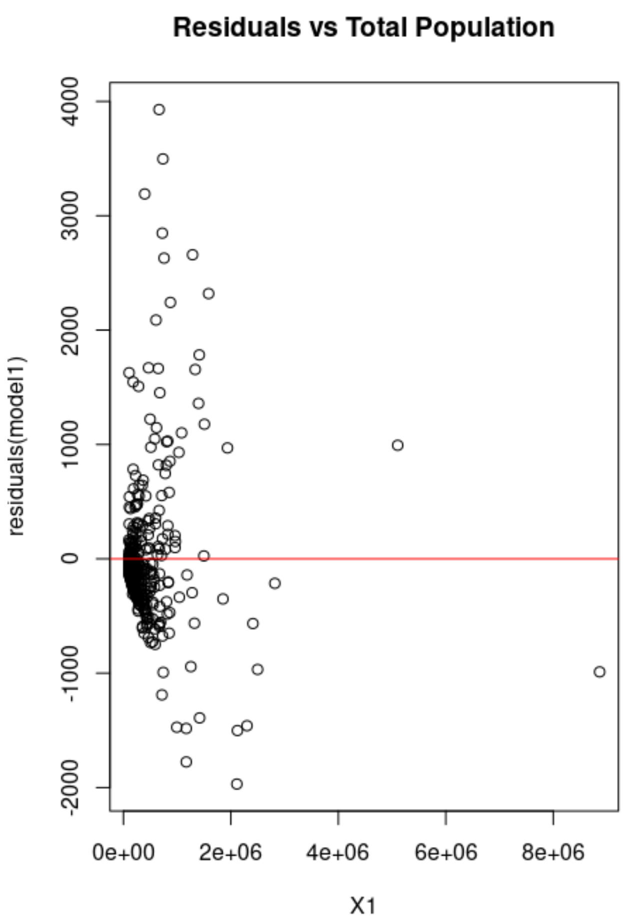
```r
# Question 3.25
# Load the necessary libraries
library(faraway)
library(car)
library(Hmisc)
library(psych)

# Loading the data
data <- read.table("https://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData
/Appendix%20C%20Data%20Sets/APPENC02.txt", header=FALSE)

# Extracting relevant columns
Y <- data$V8
X1 <- data$V5  # total population
X2 <- data$V9  # number of hospital beds
X3 <- data$V16 # total personal income

# Fitting the model for total population
model1 <- lm(Y ~ X1)
summary(model1)
# Residuals vs. total population
par(mfrow=c(1,2))
plot(X1, residuals(model1), main="Residuals vs Total Population")
abline(h=0, col="red")

# QQ-plot for residuals
qqnorm(residuals(model1), main="QQ-Plot of residuals for Total Population")
qqline(residuals(model1))

# Fitting the model for number of hospital beds
model2 <- lm(Y ~ X2)
summary(model2)
# Residuals vs. number of hospital beds
plot(X2, residuals(model2), main="Residuals vs Number of Hospital Beds")
abline(h=0, col="red")

# QQ-plot for residuals
qqnorm(residuals(model2), main="QQ-Plot of residuals for Number of Hospital Beds")
qqline(residuals(model2))

# Fitting the model for total personal income
model3 <- lm(Y ~ X3)
summary(model3)
# Residuals vs. total personal income
plot(X3, residuals(model3), main="Residuals vs Total Personal Income")
abline(h=0, col="red")

# QQ-plot for residuals
qqnorm(residuals(model3), main="QQ-Plot of residuals for Total Personal Income")
qqline(residuals(model3))
```

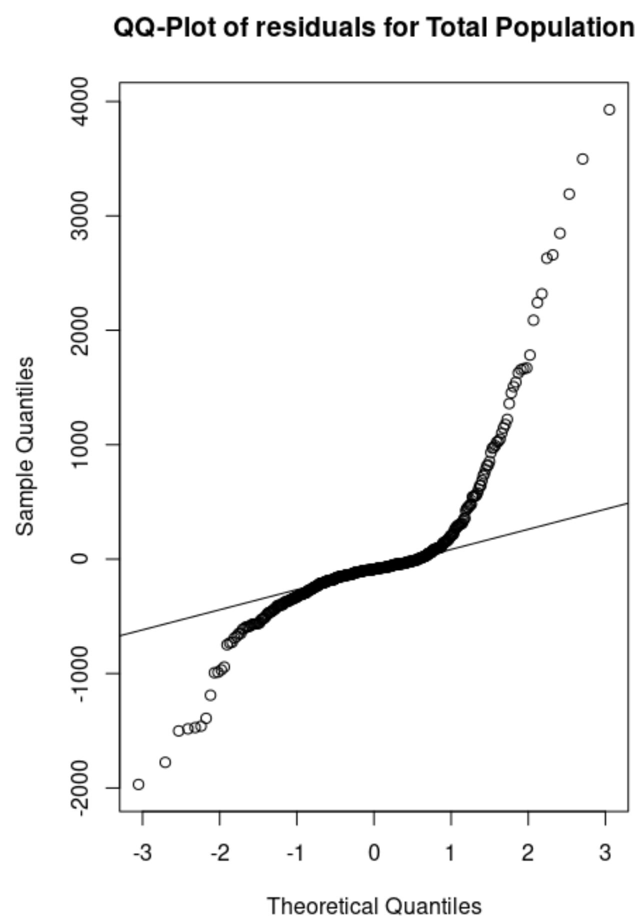
R Output:

```
> summary(model1)

Call:
lm(formula = Y ~ X1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1969.4  -209.2   -88.0    27.9  3928.7 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 ** 
X1           2.795e-03  4.837e-05  57.793  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 610.1 on 438 degrees of freedom
Multiple R-squared:  0.8841,	Adjusted R-squared:  0.8838 
F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16

> summary(model2)

Call:
lm(formula = Y ~ X2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3133.2  -216.8   -32.0    96.2  3611.1 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -95.93218   31.49396  -3.046  0.00246 ** 
X2            0.74312    0.01161  63.995  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 556.9 on 438 degrees of freedom
Multiple R-squared:  0.9034,	Adjusted R-squared:  0.9032 
F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
```

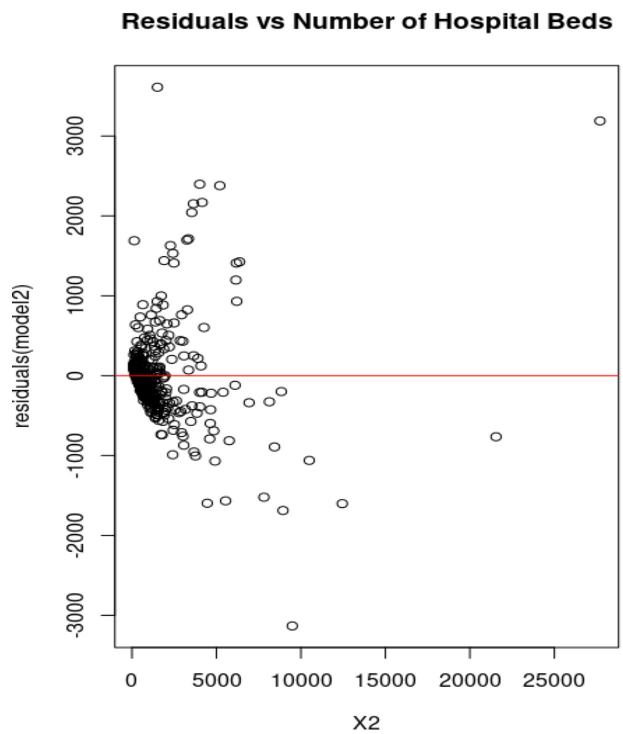
```
> summary(model3)

Call:
lm(formula = Y ~ X3)

Residuals:
    Min      1Q  Median      3Q     Max 
-1926.6  -194.5   -66.6    44.2  3819.0 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -48.39485   31.83333   -1.52    0.129    
X3            0.13170    0.00211   62.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 569.7 on 438 degrees of freedom
Multiple R-squared:  0.8989,	Adjusted R-squared:  0.8987 
F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

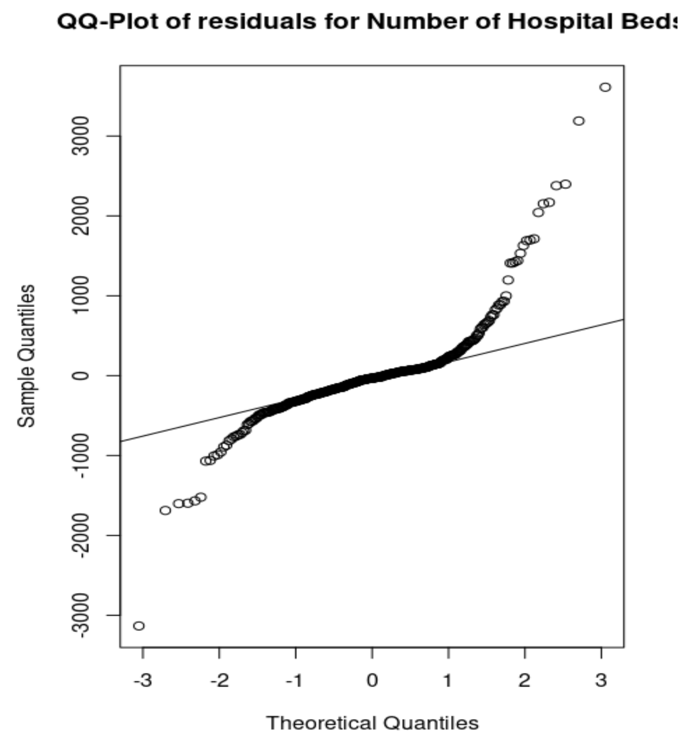
Model 1:



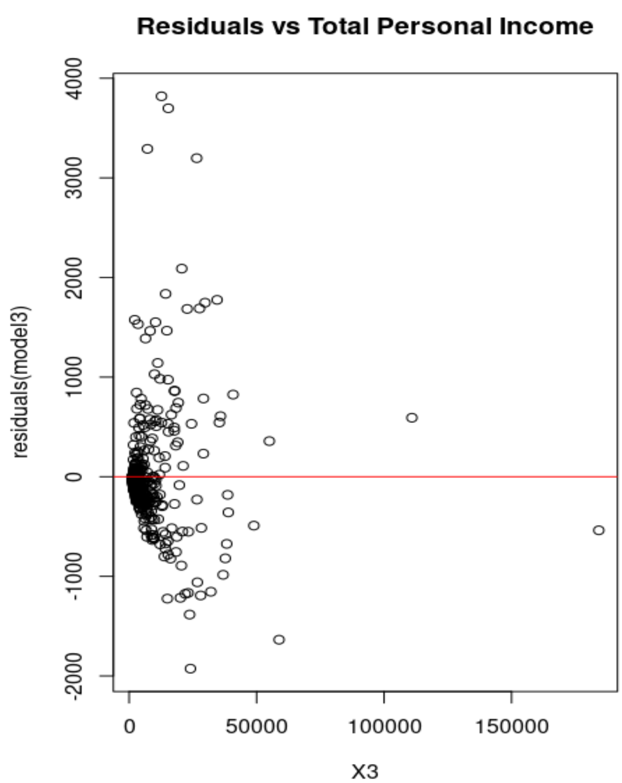



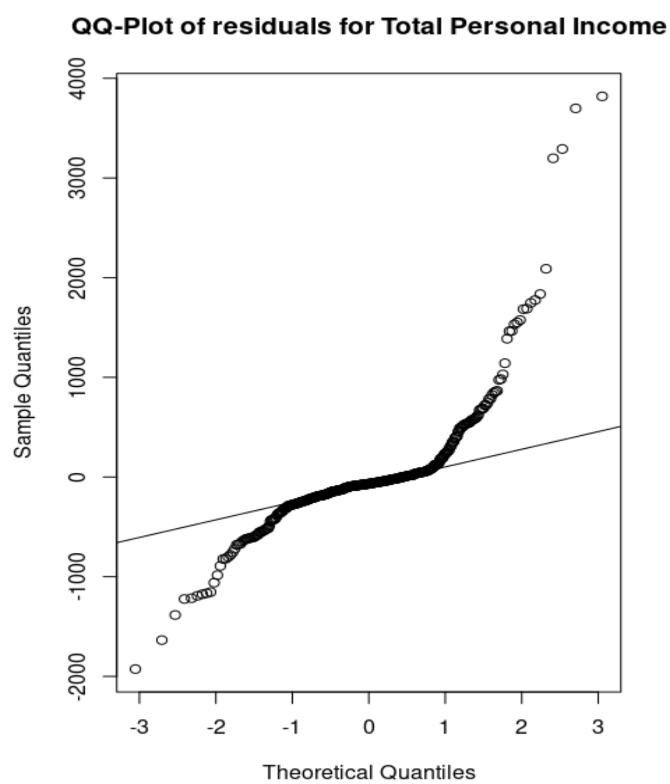
Model 2:




Model 3:



(1) Residuals vs. Predictor (or Y-hat):

- (a) Linear or Non-linear relationship: The majority of residuals are clustered on the left, and there's only a sparse representation on the right. This indicates a potential non-linear relationship between the residuals and the predictor.
- (b) Heteroskedasticity: Given the scatter of points, especially the vertically scattered points on the left, there's a potential indication of heteroskedasticity. Heteroskedasticity means the variability of the residuals is not constant across levels of the independent variable.

(2) QQ-Plot of Residuals:

- The residuals are not normally distributed as evidenced by the "N" shape of the QQ plot. This means the distribution of residuals has heavier tails than a normal distribution.

Summary:

The initial assumption for linear regression is that there's a linear relationship between predictors and the outcome, residuals are normally distributed, and there's homoskedasticity. The plots suggest potential violations of these assumptions. The non-linearity and heteroskedasticity hint at the inadequacy of a simple linear regression model for this dataset.