Spring 2023 STAT 707 Chapter 6 Homework

Reina Li

Chapter 6: 6.15, 6.16, 6.17, 6.22, the quadratic form for SS

6.15 Patient satisfaction. A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, 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 X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

```
# Import data set
df <- read.table("CH06PR15.txt")
# Set column names
colnames(df) <- c("Y", "X1", "X2", "X3")
# Data frame with interaction terms
df2 <- cbind(df, data.frame(X1X2 = df$X1 * df$X2,
                            X1X3 = df$X1 * df$X3,
                            X2X3 = df$X2 * df$X3))
```

a. Prepare a stem-and-leaf plot for each of the predictor variables. Are any noteworthy features revealed by these plots?

```
# Stem and leaf plot for X1
stem(df$X1, scale = 0.5)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   2 | 2358899999
##   3 | 01223334466678
##   4 | 0012233344557779
##   5 | 023355
```

```
# Stem and leaf plot for X2
stem(df$X2, scale = 0.25)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
```

```
##   4 | 1234666678888899999
##   5 | 00001111111122333334445678
##   6 | 02
```

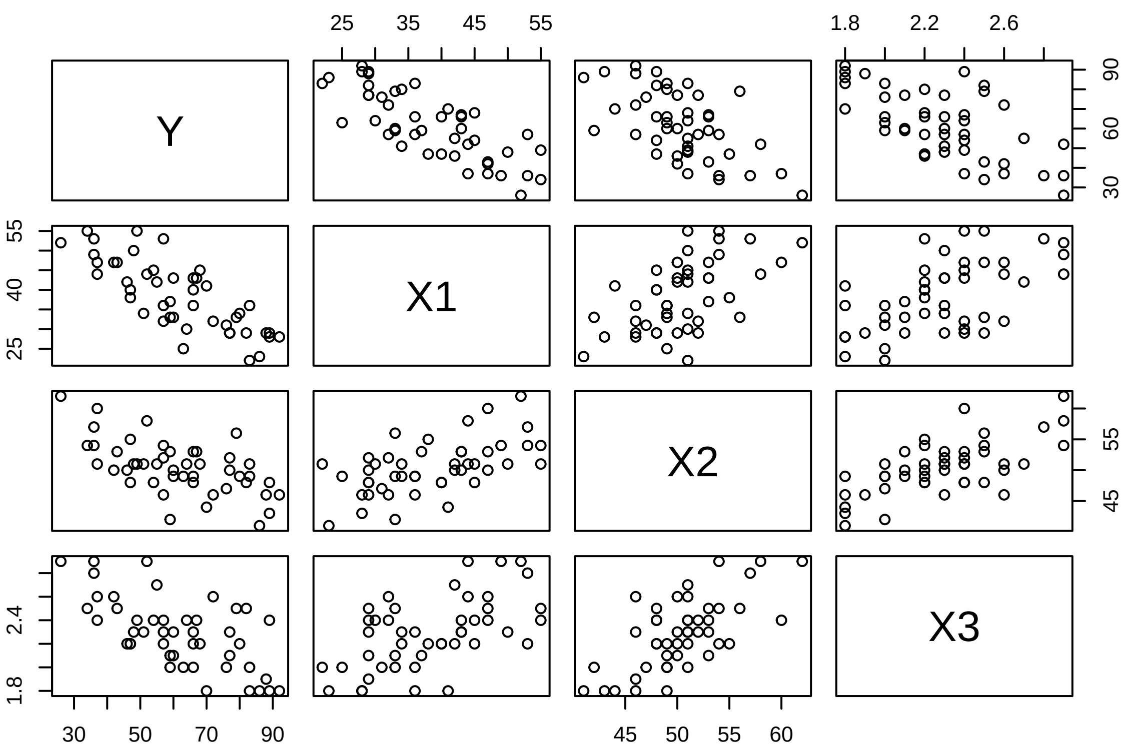
```
# Stem and leaf plot for X3
stem(df$X3, scale = 0.125)
```

```
##
##   The decimal point is at the |
##
##   1 | 888889
##   2 | 00000111222222233333344444444555566678999
```

X1 seems to be display an almost symmetric distribution with no apparent outliers. X2 seems to illustrate a right-skewed, non-normal distribution, peaking in the 50's. X3 seems to illustrate a left-skewed, non-normal distribution, peaking in the 2.0's.

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

```
# Scatter plot matrix
pairs(df)
```



```r
# Correlation matrix
cor(df)
```

```
##             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
```

From the scatter plot matrix and the correlation matrix, we can conclude the Y is negatively, linearly correlated to X_1, X_2, and X_3. The negative linear correlation is strongest between Y and X_1 (-0.787).

c. Fit regression model (6.5) for three predictor variables to the data and state the estimated regression function. How is b_2 interpreted here?

```r
# Fit regression model
model <- lm(Y ~ X1 + X2 + X3, data = df)
# View summary
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = df)
##
## 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 estimate regression function is: $\hat{Y} = 158.4913 + \text{-}1.1416 X_1 + \text{-}0.442 X_2 + \text{-}13.4702 X_3$

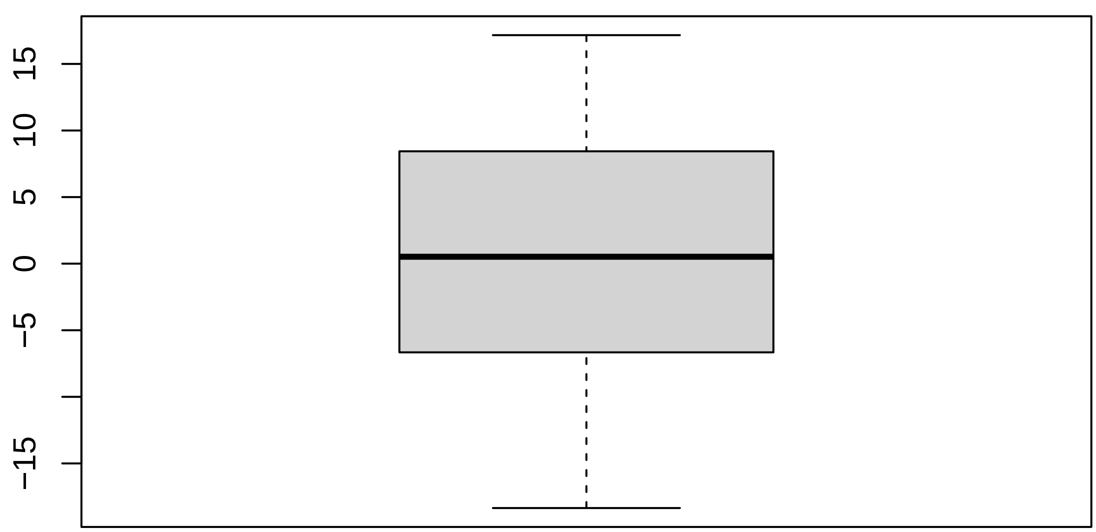
Under the assumption that the other variables X_1 and X_3 are fixed, the coefficient b_2 represents that when X_2 is increasing by 1 unit, Y will decrease by 0.4420.

d. Obtain the residuals and prepare a box plot of the residuals. Do there appear to be any outliers?

```r
# Residuals
res <- model$residuals
res
```

```
##           1           2           3           4           5           6 
##   0.1129334  -9.0796538   4.0237858   2.0093153   5.7263570  -3.6205678 
##           7           8           9          10          11          12 
## -12.8089820   0.4258777  -6.6596981   2.0030477  17.1600881  13.3526753 
##          13          14          15          16          17          18 
## -14.1654081 -15.1528562  12.5167654  -2.7946900  16.6095859   8.5409980 
##          19          20          21          22          23          24 
## -10.8725092   8.1680089   5.5810888   8.4393900   3.6796462  -3.8657107 
##          25          26          27          28          29          30 
##  -4.7338610  -4.1589620 -18.3524203   5.3949478  -9.6470593   3.3681039 
##          31          32          33          34          35          36 
## -16.3135553  11.5112774   0.6132423 -14.9762142   0.9248761  11.6161190 
##          37          38          39          40          41          42 
##  11.5071044  -5.3722872  -8.9868475  -5.7128575  11.0056590  -0.8932473 
##          43          44          45          46 
## -13.6956888  13.0578578  -5.5380448  10.0523698
```

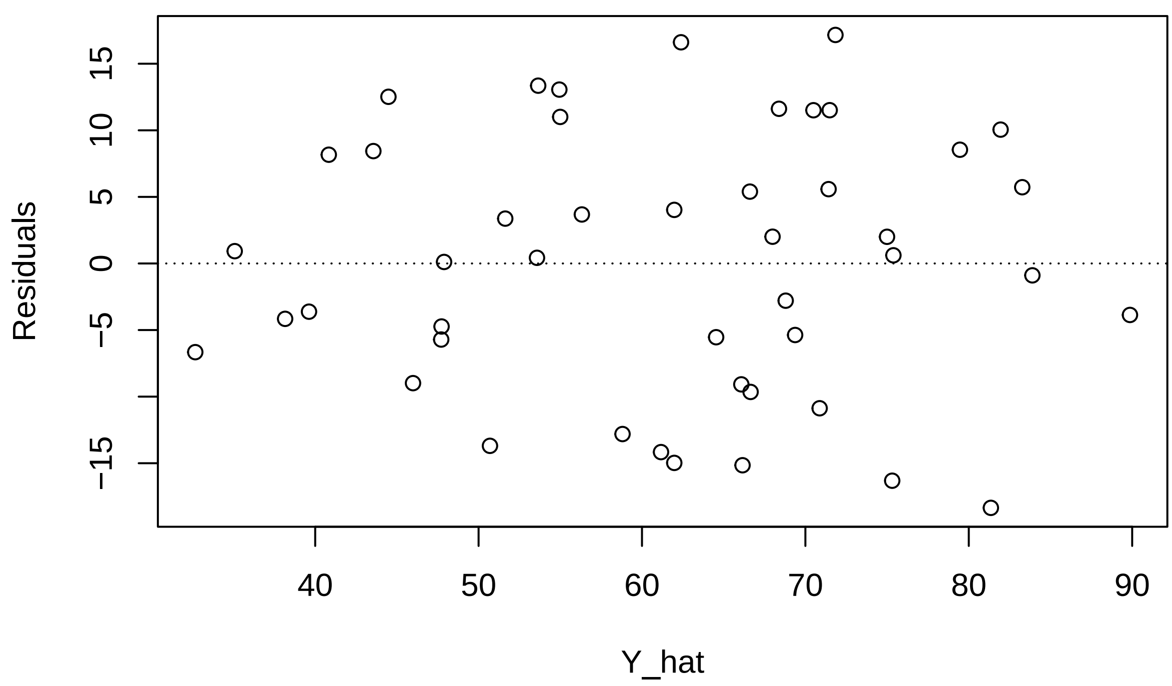
```r
# Box plot of residuals
boxplot(res)
```



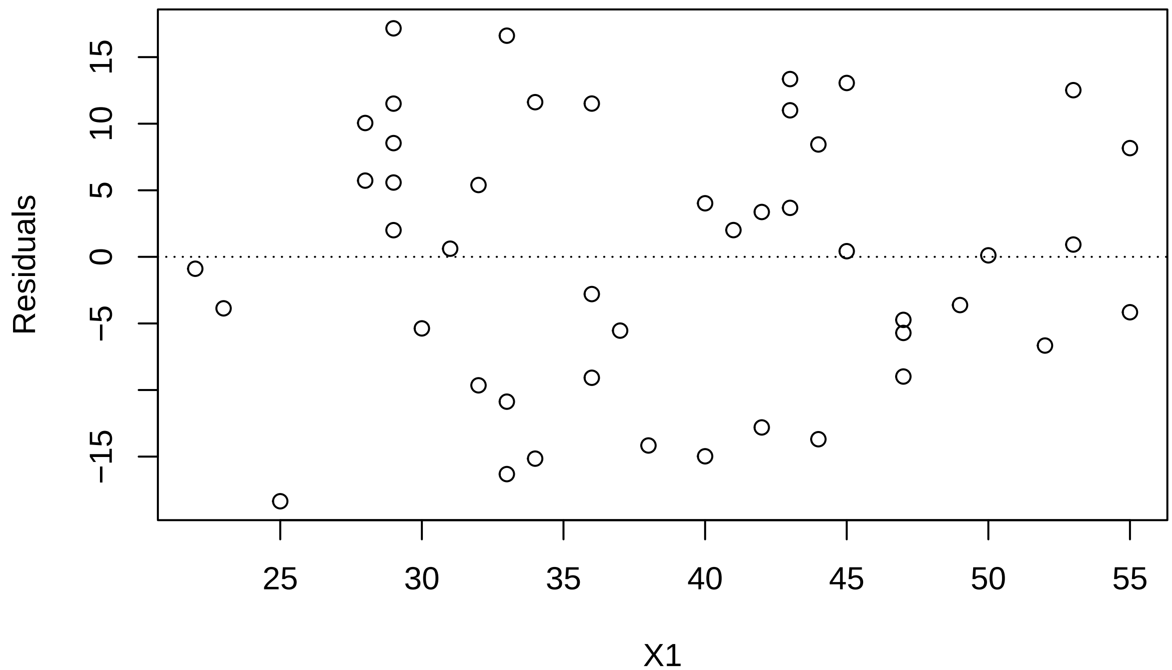
No, there does not appear to be any outliers.

e. Plot the residuals against $\hat{Y}$, each of the predictor variables, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Interpret your plots and summarize your findings.

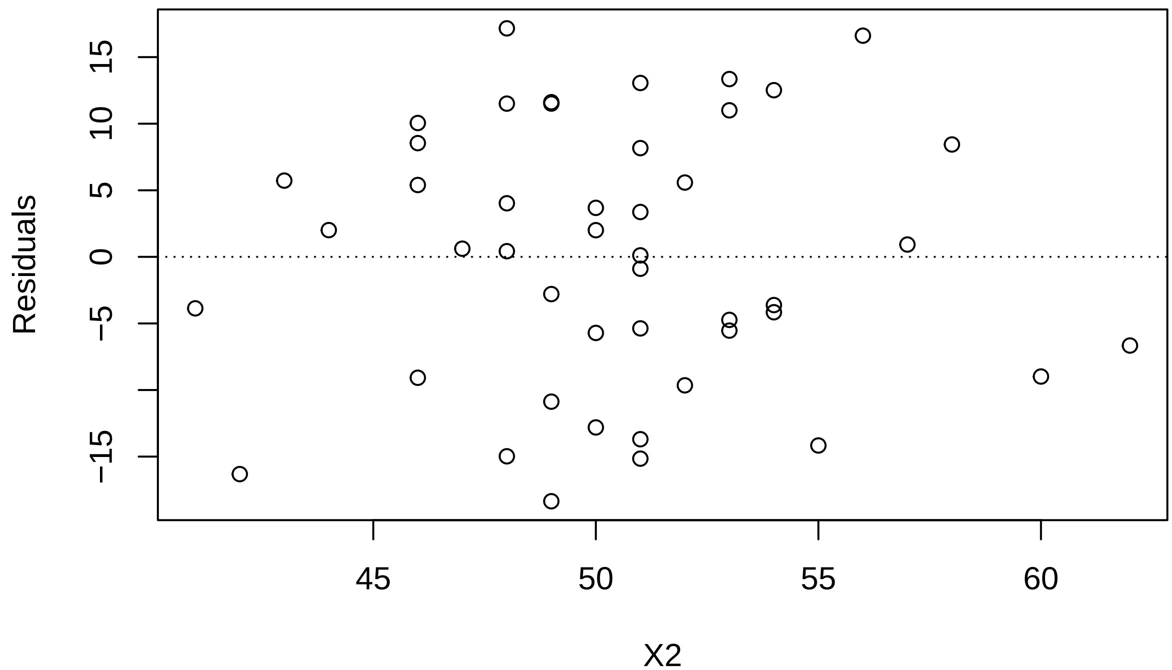
```
# Plot of residuals against Y_hat
plot(res ~ predict(model), xlab = "Y_hat", ylab = "Residuals"); abline(0,0, lty = 3)
```



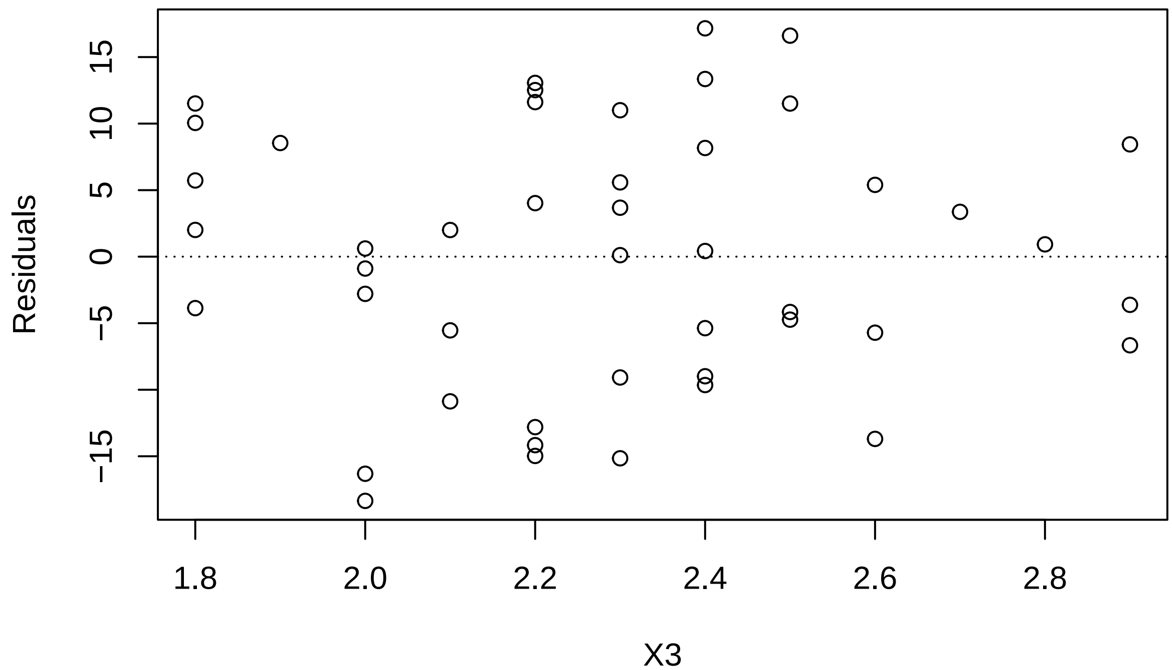

```
# Plot of residuals against X1
plot(res ~ df$X1, xlab = "X1", ylab = "Residuals"); abline(0,0, lty = 3)
```



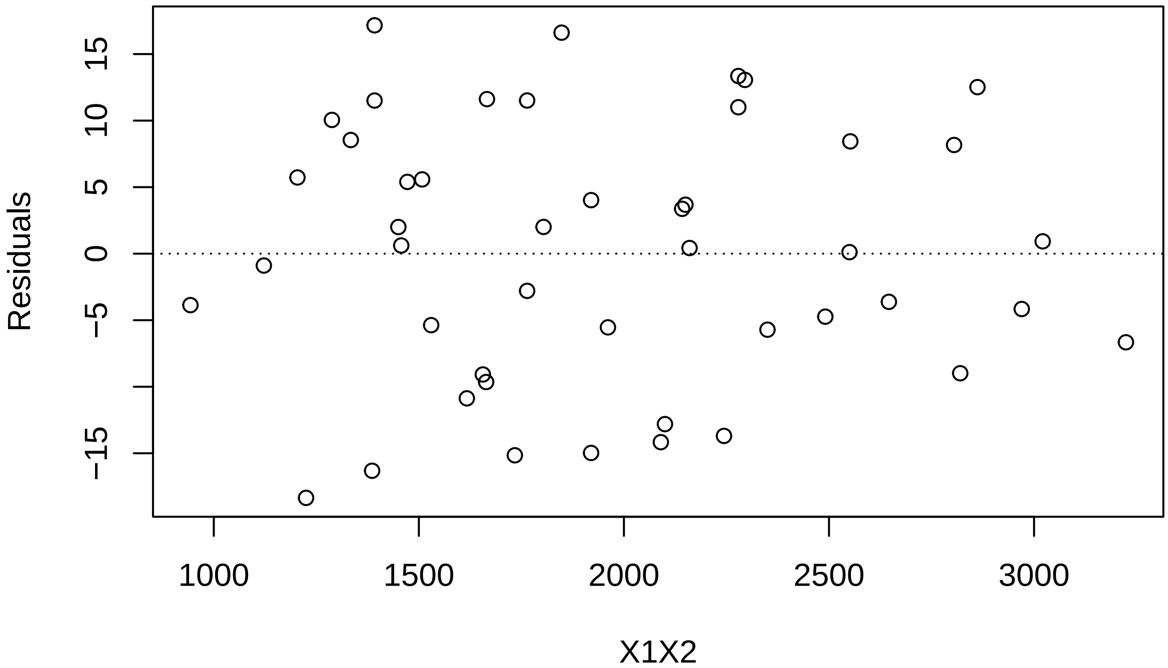

```r
# Plot of residuals against X2
plot(res ~ df$X2, xlab = "X2", ylab = "Residuals"); abline(0,0, lty = 3)
```



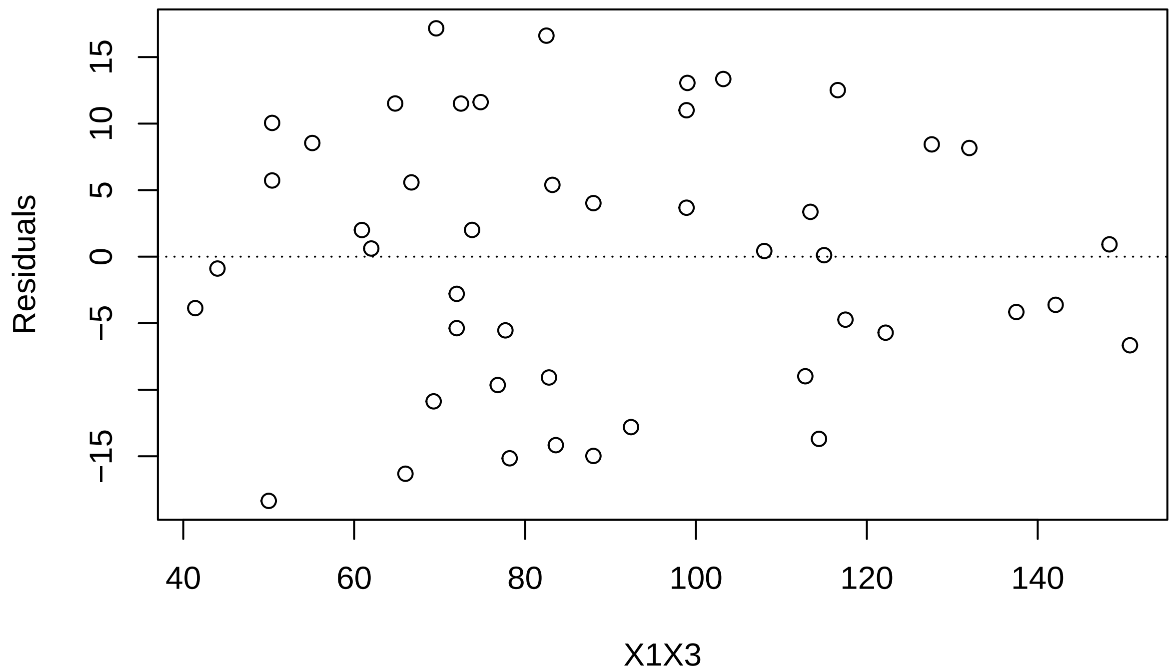

```
# Plot of residuals against X3
plot(res ~ df$X3, xlab = "X3", ylab = "Residuals"); abline(0,0, lty = 3)
```



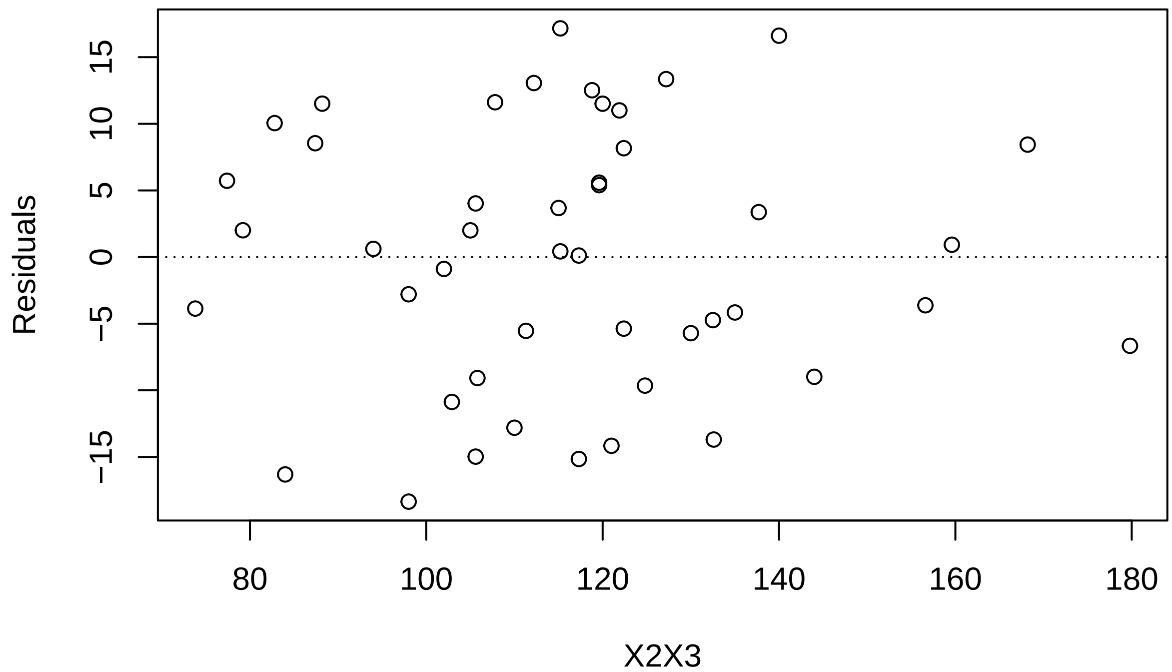

```
# Plot of residuals against X1X2
plot(res ~ df2$X1X2, xlab = "X1X2", ylab = "Residuals"); abline(0,0, lty = 3)
```



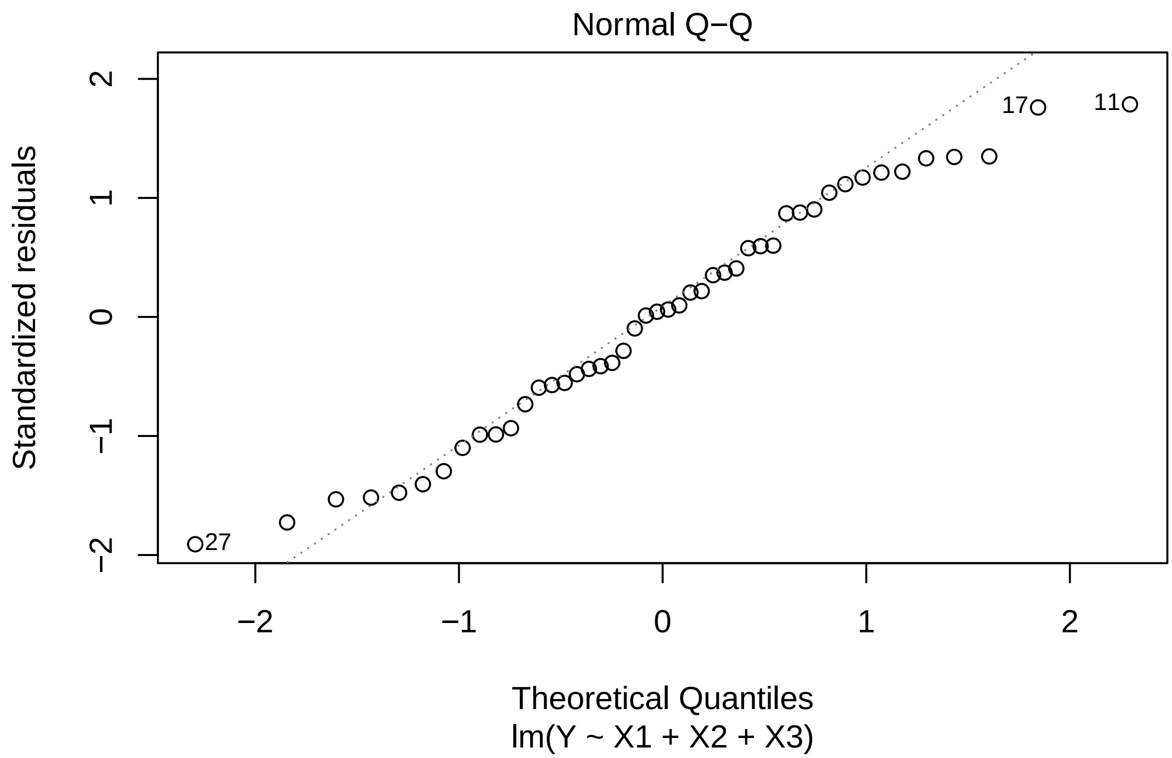

```
# Plot of residuals against X1X3
plot(res ~ df2$X1X3, xlab = "X1X3", ylab = "Residuals"); abline(0,0, lty = 3)
```




```r
# Plot of residuals against X2X3
plot(res ~ df2$X2X3, xlab = "X2X3", ylab = "Residuals"); abline(0,0, lty = 3)
```




```r
# Normal probability plot
plot(model, which = 2)
```



For the residual plots, the data points are scattered randomly around the residual = 0 line. There is no cluster and no pattern. We can conclude that a linear model is an an appropriate model.

f. Can you conduct a formal test for lack of fit here?

There is no need to conduct a formal test for lack of test because the model seems to be a good model of the data.

g. Conduct the Breusch-Pagan test for constancy of the error variance, assuming log $\sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3}$; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

```
n <- dim(df)[1]
p <- dim(df)[2]
alpha <- 0.10
X <- as.matrix(cbind(matrix(1, n, 1), df$X1, df$X2, df$X3))
Y <- as.matrix(df$Y)
b <- as.matrix(c(model$coefficients[1][[1]],
                 model$coefficients[2][[1]],
                 model$coefficients[3][[1]],
                 model$coefficients[4][[1]]))
# SSE = Y'Y - b'X'Y
sse <- t(Y) %*% Y - t(b) %*% t(X) %*% Y
sse
```

```
##           [,1]
## [1,] 4248.841
```

```r
# SSR*
# Fit a new linear regression model of squared residuals against predictor variables
model_ssr <- lm(res^2 ~ X1 + X2 + X3, data = df)
# Extract the SSR*
ssr_star <- anova(model_ssr)$`Sum Sq`[1]
ssr_star
```

```
## [1] 15084.85
```

```r
# Test statistic: X_(BP)^2* = SSR*/2 / (SSE/n)^2
test_stat <- (ssr_star/2) / ((sse/n)^2)
test_stat
```

```
##            [,1]
## [1,] 0.8840685
```

```r
# Critical value
crit_val <- qchisq(1-0.01,p-1)
crit_val
```

```
## [1] 11.34487
```

$H_0 : \gamma_1 = 0, \gamma_2 = 0$, and $\gamma_3 = 0$ (error variance is constant/homoskedasticity)
H_α : at least one $\gamma_k \neq 0$, $(k = 1, 2, 3)$ (error variance is not constant/heteroskedasticity)

$X^2_{BP} = 0.8841$
$\chi^2_{(0.99,3)} = 11.3449$

Decision rule:
If $X^2_{BP} \leq \chi^2_{(0.99,3)}$, conclude H_0
If $X^2_{BP} > \chi^2_{(0.99,3)}$, conclude H_α

Conclusion:
$X^2_{BP} \leq \chi^2_{(0.99,3)}$
Conclude H_0. Fail to reject the null hypothesis H_0. The test implies that $\gamma_1 = 0, \gamma_2 = 0$, and $\gamma_3 = 0$ and that the error variance is constant.

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 β_1, β_2, and β_3? What is the P-value of the test?

```r
# SSTO = Y'Y - (1/n)Y'JY
ssto <- t(Y) %*% Y - (1/n) * t(Y) %*% matrix(1,n,n) %*% Y
ssto
```

```
##          [,1]
## [1,] 13369.3
```

```r
# SSE = Y'Y - b'X'Y
sse <- t(Y) %*% Y - t(b) %*% t(X) %*% Y
sse
```

```
##           [,1]
## [1,] 4248.841
```

```r
# SSR = SSTO - SSE
ssr <- ssto - sse
ssr
```

```
##           [,1]
## [1,] 9120.464
```

```r
# MSR = SSR / p-1
msr <- ssr / (p-1)
msr
```

```
##           [,1]
## [1,] 3040.155
```

```r
# MSE = SSE / n-p
mse <- sse / (n-p)
mse
```

```
##           [,1]
## [1,] 101.1629
```

```r
# Test statistic: F* = MSR / MSE
# Alternative method: summary(model)$fstatistic[1]
test_stat <- msr / mse
test_stat
```

```
##           [,1]
## [1,] 30.05208
```

```r
# Critical value
crit_val <- qf(1-alpha, p-1, n-p)
crit_val
```

```
## [1] 2.219059
```

```r
# P-value
p_val <- 1 - pf(test_stat, p-1, n-p)
p_val
```

```
##               [,1]
## [1,] 1.541973e-10
```

$H_0 : \beta_1 = 0, \beta_2 = 0$, and $\beta_3 = 0$
H_α : at least one $\beta_k \neq 0$, $(k = 1, 2, 3)$

F* = 30.0521
F(0.90, 3, 42) = 2.2191

Decision rule:
If F* $\leq$ F(0.90, 3, 42), conclude H_0
If F* $>$ F(0.90, 3, 42), conclude H_α

Conclusion:
F* $>$ F(0.90, 3, 42)
Conclude H_α. Reject H_0. The test implies that at least one of $\beta_k \neq 0, (k = 1, 2, 3)$ and that Y is related to X_1, X_2, and X_3.

P-value: $1.5419732 \times 10^{-10}$

b. Obtain joint interval estimates of β_1, β_2, and β_3, using a 90 percent family confidence coefficient. Interpret your results.

```
# b_k +- B_s{b_k}
# where B = t(1-alpha/2g; n-p)
# g: # of parameters
s_squared_b <- as.numeric(mse) * solve(t(X) %*% X)
s_b1 <- s_squared_b[2,2]^0.5
s_b2 <- s_squared_b[3,3]^0.5
s_b3 <- s_squared_b[4,4]^0.5
B <- qt((1-0.10/(2*3)), n-p)
# Confidence interval for beta_1
beta1_lwr <- b[2] - B * s_b1
beta1_upr <- b[2] + B * s_b1
beta1_int <- c(beta1_lwr, beta1_upr)
names(beta1_int) <- c("lower", "upper")
beta1_int
```

```
##      lower      upper
## -1.6142482 -0.6689755
```

```
# Confidence interval for beta_2
beta2_lwr <- b[3] - B * s_b2
beta2_upr <- b[3] + B * s_b2
beta2_int <- c(beta2_lwr, beta2_upr)
names(beta2_int) <- c("lower", "upper")
beta2_int
```

```
##      lower      upper
## -1.5245098  0.6405013
```

```
# Confidence interval for beta_3
beta3_lwr <- b[4] - B * s_b3
beta3_upr <- b[4] + B * s_b3
beta3_int <- c(beta3_lwr, beta3_upr)
names(beta3_int) <- c("lower", "upper")
beta3_int
```

```
##      lower      upper
## -29.092028   2.151701
```

There is 90% confidence that the true β_1 will be between -1.6142 and -0.669. There is 90% confidence that the true β_2 will be between -1.5245 and 0.6405. There is 90% confidence that the true β_3 will be between -29.092 and 2.1517.

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

```r
# R^2 = SSR / SSTO
R2 <- ssr / ssto
R2
```

```
##           [,1]
## [1,] 0.6821943
```

It indicates that when the three predictor values (X_1, X_2, and X_3) are considered, the variation in Y is reduced by 68.22%.

6.17 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. Obtain an interval estimate of the mean satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Use a 90 percent confidence coefficient. Interpret your confidence interval.

```r
predict_int <- predict(model, data.frame(X1 = 35, X2 = 45, X3 = 2.2),
                       interval = "confidence", level = 0.90)
predict_int
```

```
##        fit      lwr      upr
## 1 69.01029 64.52854 73.49204
```

There is 90% confidence that the true response value is between 64.5285 and 73.492.

b. Obtain a predictor interval for a new patient's satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Use a 90 percent confidence coefficient. Interpret your prediction interval.

```r
predict_int <- predict(model, data.frame(X1 = 35, X2 = 45, X3 = 2.2),
                       interval = "prediction", level = 0.90)
predict_int
```

```
##        fit      lwr      upr
## 1 69.01029 51.50965 86.51092
```

There is 90% confidence that the true response value is between 51.5097 and 86.5109.

6.22 For each of the following regression models, indicate whether it is a general linear regression model. If it is not, state whether it can be expressed in the form of (6.7) by a suitable transformation:

a. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 log_{10} X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$

Yes, this is a general linear regression model.

b. $Y_i = \varepsilon_i exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)$

No, this is not a general linear regression model. That is because there are non-linear predictor terms. It can be expressed in the form of (6.7) by taking the ln of Y_i:
$Y_i = \varepsilon_i exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)$
$ln(Y_i) = ln(\varepsilon_i exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2))$
$ln(Y_i) = ln(\varepsilon_i) + ln(exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2))$
$ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2 + ln(\varepsilon_i)$

c. $Y_i = log_{10}(\beta_1 X_{i1}) + \beta_2 X_{i2} + \varepsilon_i$

Yes, this is a general linear regression model.

d. $Y_i = \beta_0 exp(\beta_1 X_{i1}) + \varepsilon_i$

No, this is not a general linear regression model. That is because it contains an exponential predictor term. It can be expressed in the form of (6.7) by taking the ln of Y_i:
$Y_i = \beta_0 exp(\beta_1 X_{i1}) + \varepsilon_i$
$ln(Y_i) = ln(\beta_0 exp(\beta_1 X_{i1}) + \varepsilon_i)$
$ln(Y_i) = ln(\beta_0) + ln(exp(\beta_1 X_{i1}) + ln(\varepsilon_i))$
$ln(Y_i) = ln(\beta_0) + \beta_1 X_{i1} + ln(\varepsilon_i)$

e. $Y_i = [1 + exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)]^{-1}$

No, this is not a general linear regression model. That is because it contains exponential predictor terms. It can be expressed in the form of (6.7) by taking the ln of Y_i:
$Y_i = [1 + exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)]^{-1}$
$ln(Y_i) = ln([1 + exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)]^{-1})$
$ln(Y_i) = -ln([1 + exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i)])$
$ln(Y_i) = -ln(1) + ln(exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i))$
$ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$