HW5

2023-09-25

6.15 (a) (b) (c)
- For 6.15 (a): Instead of a stem-and-leaf plot, create a histogram for each predictor variable.
- Added question before doing 6.16: 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 iid~ N(0,σ^2)) is reasonable. **This problem is open-ended and will be graded based on efforts.** Use this opportunity to **practice your diagnostics skills and enhance your understanding.** Do as much as you can.

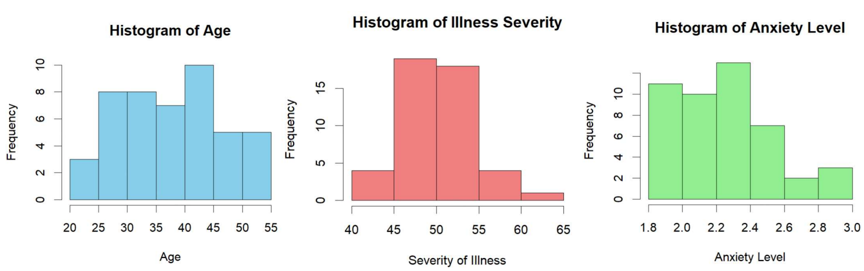
6.16 (a) (c):
- For 6.16 (a): The problem asks for the **F-test of global model utility.** Follow the **5 steps of hypothesis testing** given in the lecture slides.

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

i:	1	2	3	...	44	45	46
X_{i1}:	50	36	40	...	45	37	28
X_{i2}:	51	46	48	...	51	53	46
X_{i3}:	2.3	2.3	2.2	...	2.2	2.1	1.8
Y_i:	48	57	66	...	68	59	92

a. Create a histogram for each predictor variable. Are any noteworthy features revealed by these plots?

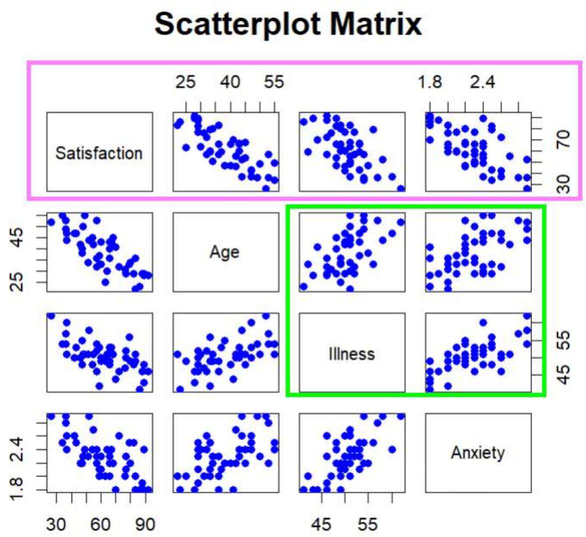
```r
############### Read data ###############

Satisfaction <- read.table("CH06PR15.txt")
head(Satisfaction)
colnames(Satisfaction) <- c("Satisfaction","Age","Illness", "Anxiety")
str(Satisfaction) #Displays a concise structure of an R object
attach(Satisfaction) #Adds a database to R's search path,
                     #allowing direct variable referencing. Use with caution.

#------------------------------------------------------------------------------------------------
############### Inspect data, histogram for each predictor variable ###############
Psych.describe(Satisfaction)
hist(Age, main="Histogram of Age", xlab="Age", col="skyblue", breaks=6)
hist(Illness, main="Histogram of Illness Severity", xlab="Severity of Illness", col="lightcoral", breaks=20)
hist(Anxiety, main="Histogram of Anxiety Level", xlab="Anxiety Level", col="lightgreen", breaks=10)
```



Histogram of Anxiety level is right skewed. Most patients have an anxiety index between 1.8 and 2.5.

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

```r
> pairs(Satisfaction[, 1:4], main="Scatterplot Matrix", pch=19, col="blue")
```



• (Pink box) Satisfaction vs. Age, illness, or Anxiety: There is a negative linear relationship, suggesting that as age, illness, or anxiety increases, satisfaction tends to decrease.

• (Green box) Among predictor variables, there are some positive relationships, between illness and Age, illness and Anxiety, and age and anxiety.

```
> cor(Satisfaction)
             Satisfaction        Age    Illness    Anxiety
Satisfaction    1.0000000 -0.7867555 -0.6029417 -0.6445910
Age            -0.7867555  1.0000000  0.5679505  0.5696775
Illness        -0.6029417  0.5679505  1.0000000  0.6705287
Anxiety        -0.6445910  0.5696775  0.6705287  1.0000000
```

• Satisfaction & Age: The correlation is -0.787, a strong negative relationship.
• Satisfaction & Severity: The correlation is -0.603, a moderate negative relationship.
• Satisfaction & Anxiety: The correlation is -0.645, a moderately strong negative relationship.
• Among the predictors, the highest correlation is observed between Severity and Anxiety with a value of 0.671, a moderately strong positive relationship.

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

Fitted model:

Satisfaction = β0 + β1*Age + β2*Illiness + β3*Anxiety

Satisfaction = 158.4913 -1.1416*Age -0.4420*Illiness -13.4702*Anxiety

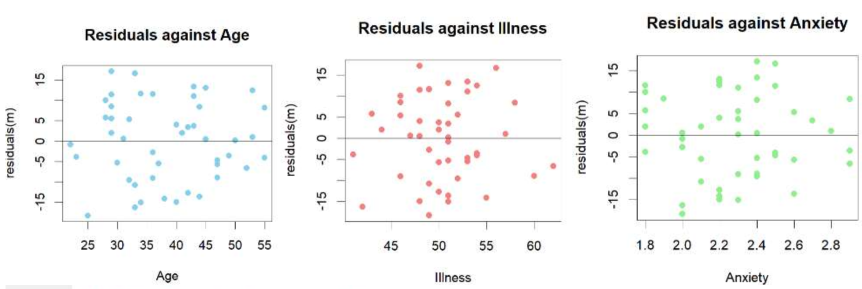
B2 interpretation: For every one-unit increase in the severity of illness, the satisfaction decreases by 0.4420 units, holding other variables constant.

```
m <- lm(Satisfaction ~ Age + Illness + Anxiety, data=Satisfaction)
summary(m)
# Satisfaction = β0 + β1*Age + β2*Illiness + β3*Anxiety
# Satisfaction = 158.4913 -1.1416*Age -0.4420*Illiness -13.4702*Anxiety

# Coefficient:
# The intercept (β0) is 158.4913, the estimated satisfaction when all predictor variables are 0.

# For every one-year increase in age, the satisfaction decreases by 1.1416 units,
#     holding other variables constant.
# For every one-unit increase in the severity of illness, the satisfaction
#     decreases by 0.4420 units, holding other variables constant.
# For every one-unit increase in anxiety, the satisfaction decreases by
#     13.4702 units, holding other variables constant.
```

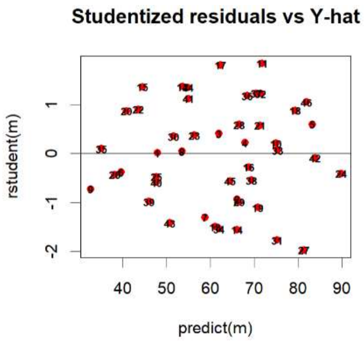
```
> ############### Fit regression model ###############
> m <- lm(Satisfaction ~ Age + Illness + Anxiety, data=Satisfaction)
> summary(m)

Call:
lm(formula = Satisfaction ~ Age + Illness + Anxiety, data = Satisfaction)

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 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Illness      -0.4420     0.4920  -0.898   0.3741    
Anxiety     -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
```

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 iid~ N(0,σ^2)) is reasonable.

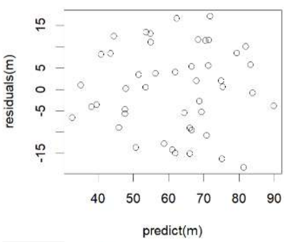
```
#-----------------------------------------------------------------------------------
############### Residual diagnostics ###############

# (1) Check whether the relationship between Y and each X is linear.
# Plot the residuals against each X.
plot(Age, residuals(m), main = "Residuals against Age", pch=19, col="skyblue")
abline(h=0)
plot(Illness, residuals(m), main = "Residuals against Illness", pch=19, col="lightcoral")
abline(h=0)
plot(Anxiety, residuals(m), main = "Residuals against Anxiety", pch=19, col="lightgreen")
abline(h=0)
# Impression: The residuals do not seem to relate to any Age, Illness,
# or Anxiety in a systematic manner.
# Thus, the first-order terms of the 3 predictors in model m seems sufficient.
```




```
# (2) Check for outliers.
# Plot the studentized residuals against Y-hat.
plot(predict(m), rstudent(m), main = "Studentized residuals vs Y-hat",
     pch=16, col="red")
text(predict(m), rstudent(m), names(rstudent(m)), cex=0.6, font=2)
abline(h=0)
outliers <- rstudent(m)[abs(rstudent(m)) > 2] #potential outlier grater than 2
outliers
# Impression: All studentized residuals are in a reasonable range given the
# approximate 1-2-3 rule.
> outliers <- rstudent(m)[abs(rstudent(m)) > 2] #potential outlier grater than 2
> outliers
named numeric(0)
```



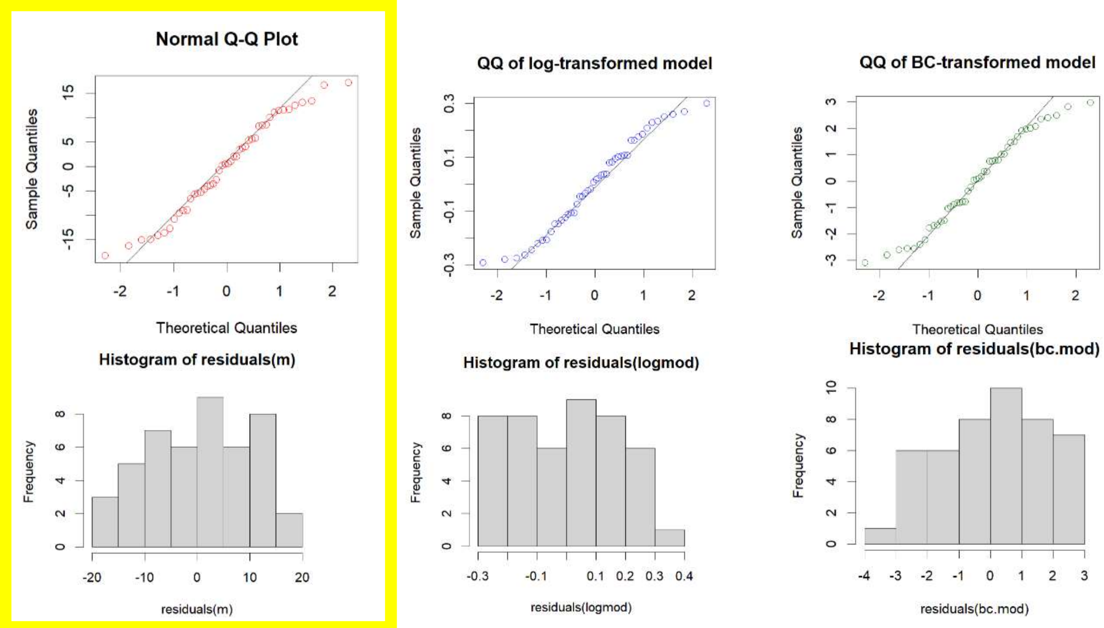

```
# (3) Check for heteroskedasticity
# Plot the residuals against Y-hat
plot(predict(m), residuals(m))
# Impression: The vertical spread of the points are roughly constant across
# different X values. No concern of keteroskedasticity.
```




```
# Shapiro test.
# ncvTest() function from the car package performs a non-constant variance score test
# (known as the Breusch-Pagan test) to test for heteroscedasticity in a linear regression model.
# H0: the variances of the residuals are constant (homoscedasticity),
# H1: they are not constant (heteroscedasticity).
ncvTest(logmod)
ncvTest(bc.mod)
ncvTest(m)
# impression: p=value for all 3 transformed models are high.
# There's no evidence to reject H0, meaning homoscedastic.
```

```
> # Shapiro test.
> # ncvTest() function from the car package performs a non-constant variance score test
> # (known as the Breusch-Pagan test) to test for heteroscedasticity in a linear regression model.
> # H0: the variances of the residuals are constant (homoscedasticity),
> # H1: they are not constant (heteroscedasticity).
> ncvTest(logmod)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.9292462, Df = 1, p = 0.33506
> ncvTest(bc.mod)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.08271333, Df = 1, p = 0.77365
> ncvTest(m)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.6513299, Df = 1, p = 0.41964
```

```
# (4) Check whether the residuals are normally distributed.
# Create the QQ-plot of the residuals.
qqnorm(residuals(m), col="red")
qqline(residuals(m)) # Our plot shows a slight deviation in the tails but
#                      is mostly aligned with the line, suggesting
#                      the residuals are approximately normally distributed.
hist(residuals(m), breaks=10)
```




```
# (5) whether the observations are independent from each other
# Each row in the data represents a patient.
# Also, residuals fluctuate in a more or less random pattern around the baseline 0.
# Thus, it is reasonable to assume the patients are independent with respect to
# the three measures involved in this regression problem.
```

Summary: the assumptions of the normal error regression (NER) model (i.e., ε iid~ N(0,σ^2)) is reasonable.

1) The relationship between Y and each X is linear,
2) No outlier
3) Homoscedastic: The variance of the errors is constant across all levels of the independent variables.
4) ε (0, σ^2): slightly deviated in the tails but is mostly aligned with the line, suggesting the residuals are approximately normally distributed.
5) independent observations: patients are independent

***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 α = 0.10.

The problem asks for the **F-test of global model utility.** Follow the **5 steps of hypothesis testing** given in the lecture slides.

Step 1. Check assumption.

$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

The complet diagnostic is shown above.

(1) Linear relationship between X and Y.

(2) No outlier is shown

(3) Homosketastic

(4) NER. residuals are namally distributed.

(5) Independent observation

Step 2. Hypothesis.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

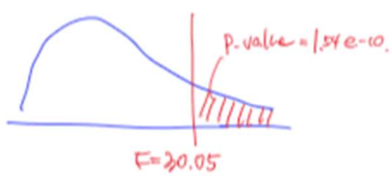
H_1 : At least one of the regression coefficients is Nonzero.
(There exists linear association between Y and at least one of predictors.

Step 3. Find test statistic

$$F_{obs} = \frac{MSR}{MSE} = \frac{3040.155}{101.2} = 30.05$$

- $MSR = SSR/P-1 \quad (P=4)$
  $= 9120.5/4$ ——— *
  $= 3040.155$
- $MSE = 101.2$ ——— *

Step 4. p-value.



1 − pf (q= 30.05, df1 = 4-1, df2 = 46-4)
(= pf(q = 30.05, df1=4-1, df2=46-4, lower.tail=F)
= 1.543485 e-10

Step 5. conclusion.

Because p-value < 0.1 = α, we reject H_0.
There is enough evidence suggesting that the true population slop is not zero.
Thus, we conclude that there is linear association between Y and X (at least one of Xs)

```
> # Find the p-value of the F-test mannually
> 1 - pf(q=30.05, df1=4-1 , df2=46-4)
[1] 1.543485e-10
```

```
> # Manually calculate the elements in the ANOVA table
> Y <- Satisfaction$Satisfaction
> Y.bar <- mean(Y)
> Y.hat <- predict(m)

> SSTO <- sum((Y - Y.bar) ^ 2)
> SSR <- sum((Y.hat - Y.bar) ^ 2)
> SSE <- sum((Y - Y.hat) ^ 2)
> n <- nrow(Satisfaction)
> p <- 4
> df.TO <- n - 1
> df.R <- p - 1
> df.E <- n - p
> MSTO <- SSTO / df.TO
> MSR <- SSR / df.R
> MSE <- SSE / df.E
> SSTO; SSR; SSE
[1] 13369.3
[1] 9120.464
[1] 4248.841
> df.TO; df.R; df.E
[1] 45
[1] 3
[1] 42
> MSTO; MSR; MSE
[1] 297.0957
[1] 3040.155
[1] 101.1629
```

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

R^2 = SSR/SST = 1 – SSE/SST

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

Response: Satisfaction
          Df Sum Sq Mean Sq F value    Pr(>F)    
Age        1 8275.4  8275.4 81.8026 2.059e-11 ***
Illness    1  480.9   480.9  4.7539   0.03489 *  
Anxiety    1  364.2   364.2  3.5997   0.06468 .  
Residuals 42 4248.8   101.2                      
```

- **SSE= 4248.8**
- **SSR = 8275.4 + 480.9 + 364.2 = 9120.5**
- **SST = 4248.8 + 9120.5 = 13369.3**

- **Thus, R^2= 9120.5/13369.3 = 0.6821972 , this is equal to R-squared below.**

```
> ############### Fit regression model ###############
> m <- lm(Satisfaction ~ Age + Illness + Anxiety, data=Satisfaction)
> summary(m)

Call:
lm(formula = Satisfaction ~ Age + Illness + Anxiety, data = Satisfaction)

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 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Illness      -0.4420     0.4920  -0.898   0.3741    
Anxiety     -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
> sqrt(.6822)
[1] 0.825954
```

68.22% (R^2) of the variance in satisfaction is explained by the model.
The multiple correlation coefficient r for the regression model is 0.825954 (sqrt 0.6822), indicating a reasonably strong correlation between the dependent variable (Satisfaction) and the combined predictor variables (Age, illness, anxiety). The model is statistically significant with p=1.542e-10 for the F-statistic (30.05).

R code:

Load packages

```
rm(list=ls()) # Clean up the workspace for the new analysis

# Set the following to your own folder
setwd("C:/Users/jyang/OneDrive - Arizona State University/10 Classes_OneDrive
/2023_STP530_Regression")

#-----------------------------------------------------------------------
-------------------
############### Install and load the add-on packages ###############

# install.packages("faraway")
#     Tools and datasets for linear and generalized linear models,
#     aiding data exploration and visualization.
```

```r
# install.packages("car")
#     Utilities for regression diagnostics, hypothesis testing, & data visual
ization,
#     complementing the book "An R Companion to Applied Regression.

# install.packages("Hmisc")
#     A suite for data analysis, especially in clinical research,
#     offering data imputation, summarization, and plotting.

# install.packages("psych")
#     Designed for psychometric research, it provides tools
#     for factor analysis, reliability analysis, and data visualization.

# install.packages("rgl")
#     Enables interactive 3D visualizations using OpenGL,
#     suitable for visualizing complex datasets and shapes.


# Load packages: Every time opining a new R session, need to load packages.
library(faraway)

## Warning: package 'faraway' was built under R version 4.3.1

library(car)

## Warning: package 'car' was built under R version 4.3.1

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.1

##
## Attaching package: 'car'

## The following objects are masked from 'package:faraway':
##
##     logit, vif

library(Hmisc) # describe functions of Hmisc and Psych are same, thus rename
it.

## Warning: package 'Hmisc' was built under R version 4.3.1

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

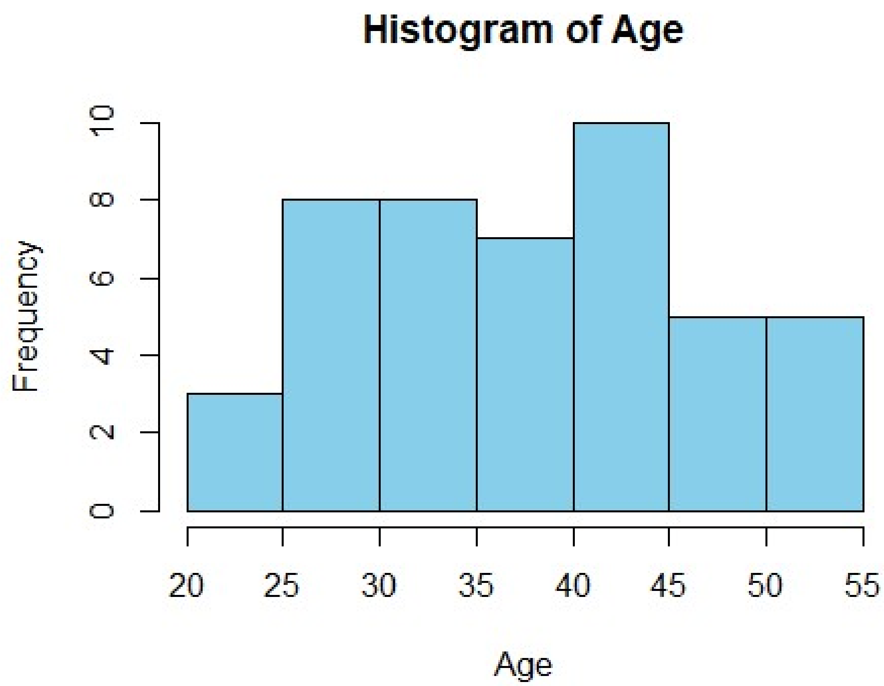
```
Hmisc.describe <- describe
library(psych) # describe functions of Hmisc and Psych are same, thus rename
it.

## Warning: package 'psych' was built under R version 4.3.1

##
## Attaching package: 'psych'

## The following object is masked from 'package:Hmisc':
##
##     describe

## The following object is masked from 'package:car':
##
##     logit

## The following object is masked from 'package:faraway':
##
##     logit

Psych.describe <- describe
library(rgl)

## Warning: package 'rgl' was built under R version 4.3.1
```

Read data, inspect data

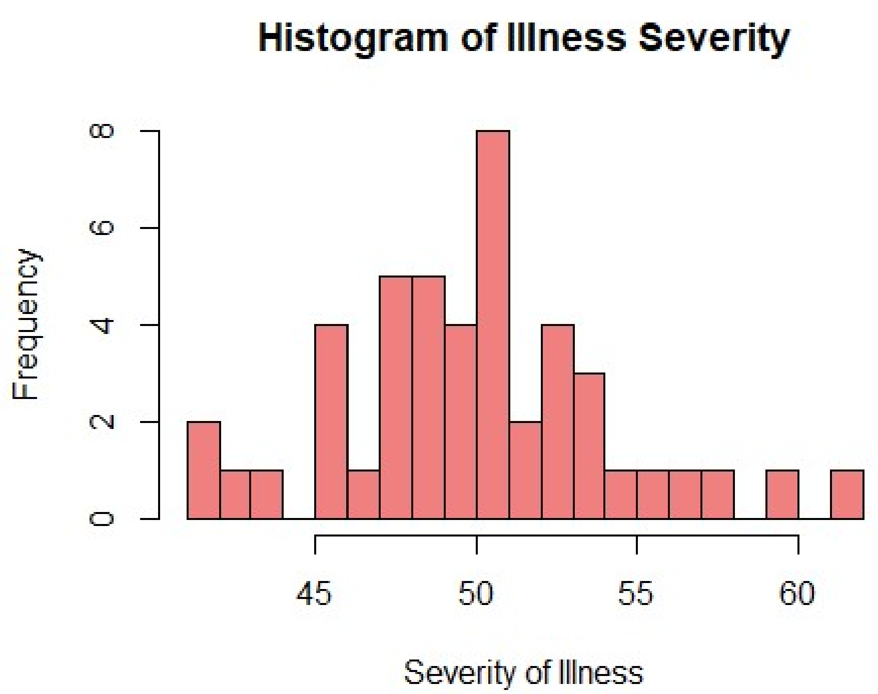
6.16a. Histogram for each predictor variable

```
############### Read data ###############

Satisfaction <- read.table("CH06PR15.txt")
head(Satisfaction)

##   V1 V2 V3  V4
## 1 48 50 51 2.3
## 2 57 36 46 2.3
## 3 66 40 48 2.2
## 4 70 41 44 1.8
## 5 89 28 43 1.8
## 6 36 49 54 2.9

colnames(Satisfaction) <- c("Satisfaction","Age","Illness", "Anxiety")
str(Satisfaction) #Displays a concise structure of an R object

## 'data.frame':    46 obs. of  4 variables:
##  $ Satisfaction: int  48 57 66 70 89 36 46 54 26 77 ...
##  $ Age         : int  50 36 40 41 28 49 42 45 52 29 ...
##  $ Illness     : int  51 46 48 44 43 54 50 48 62 50 ...
##  $ Anxiety     : num  2.3 2.3 2.2 1.8 1.8 2.9 2.2 2.4 2.9 2.1 ...

attach(Satisfaction) #Adds a database to R's search path,
```

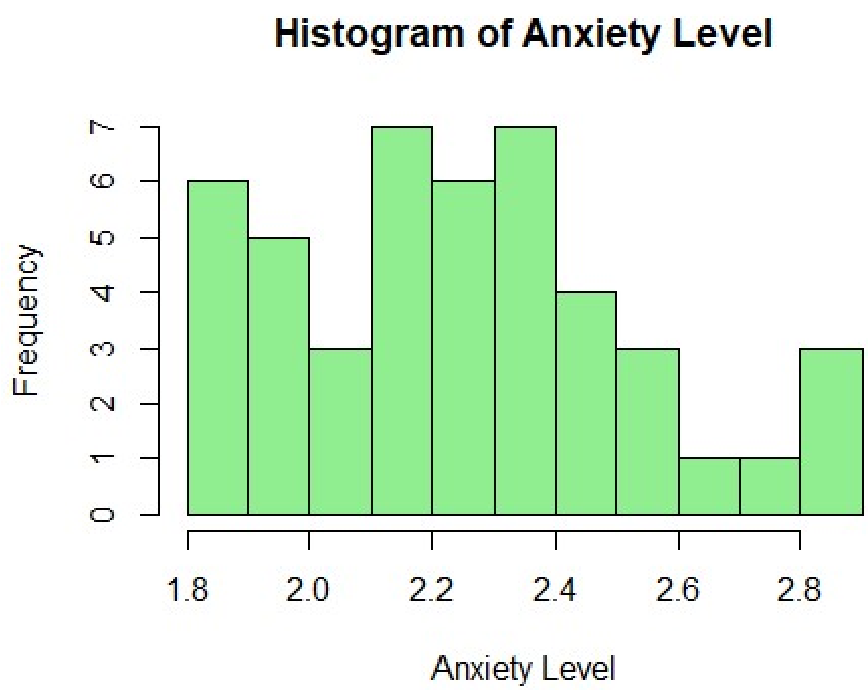
```
## The following object is masked _by_ .GlobalEnv:
##
##     Satisfaction

                     #allowing direct variable referencing. Use with caution.

#-----------------------------------------------------------------------------
-------------------
############### Inspect data, histogram for each predictor variable #########
######
Psych.describe(Satisfaction)

##              vars  n  mean    sd median trimmed   mad  min  max range  ske
w
## Satisfaction    1 46 61.57 17.24   60.0   61.63 19.27 26.0 92.0  66.0 -0.0
2
## Age             2 46 38.39  8.92   37.5   38.21 10.38 22.0 55.0  33.0  0.1
5
## Illness         3 46 50.43  4.31   50.5   50.34  3.71 41.0 62.0  21.0  0.2
8
## Anxiety         4 46  2.29  0.30    2.3    2.28  0.30  1.8  2.9   1.1  0.2
1
##              kurtosis   se
## Satisfaction    -1.01 2.54
## Age             -1.04 1.31
## Illness          0.33 0.64
## Anxiety         -0.56 0.04

hist(Age, main="Histogram of Age", xlab="Age", col="skyblue", breaks=6)
```




```
hist(Illness, main="Histogram of Illness Severity", xlab="Severity of Illness
", col="lightcoral", breaks=20)
```

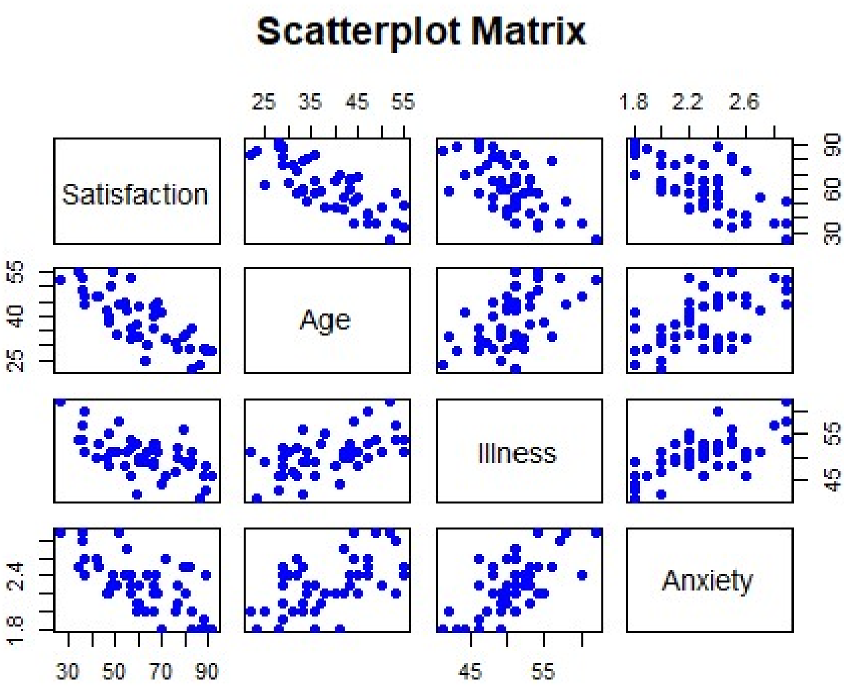


```
hist(Anxiety, main="Histogram of Anxiety Level", xlab="Anxiety Level", col="l
ightgreen", breaks=10)
```



6.15b. scatter plot matrix and the correlation matrix

```
############### scatter plot matrix and the correlation matrix ##############
#
pairs(Satisfaction[, 1:4], main="Scatterplot Matrix", pch=19, col="blue")
```




```
# • Satisfaction vs. Age: there's a negative linear relationship,
#       as age increases, satisfaction tends to decrease.
# • Satisfaction vs. Severity: there's a negative trend,
#       as the severity of illness increases, satisfaction might decrease.
# • Satisfaction vs. Anxiety: A negative trend is observed,
#       a possible decrease in satisfaction with increasing anxiety levels.
# • Among predictor variables, there're some positive relationships,
#       between Severity and Age, Severity and Anxiety.
cor(Satisfaction)

##              Satisfaction        Age    Illness    Anxiety
## Satisfaction    1.0000000 -0.7867555 -0.6029417 -0.6445910
## Age            -0.7867555  1.0000000  0.5679505  0.5696775
## Illness        -0.6029417  0.5679505  1.0000000  0.6705287
## Anxiety        -0.6445910  0.5696775  0.6705287  1.0000000

# • Satisfaction & Age: The correlation is -0.787,
#     a strong negative relationship.
# • Satisfaction & Severity: The correlation is -0.603,
#     a moderate negative relationship.
# • Satisfaction &d Anxiety: The correlation is -0.645,
#     a moderately strong negative relationship.
# • Among the predictors, the highest correlation is between Severity & Anxie
ty
#     with a value of 0.671, a moderately strong positive relationship.
```

6.15c. Fit regression model

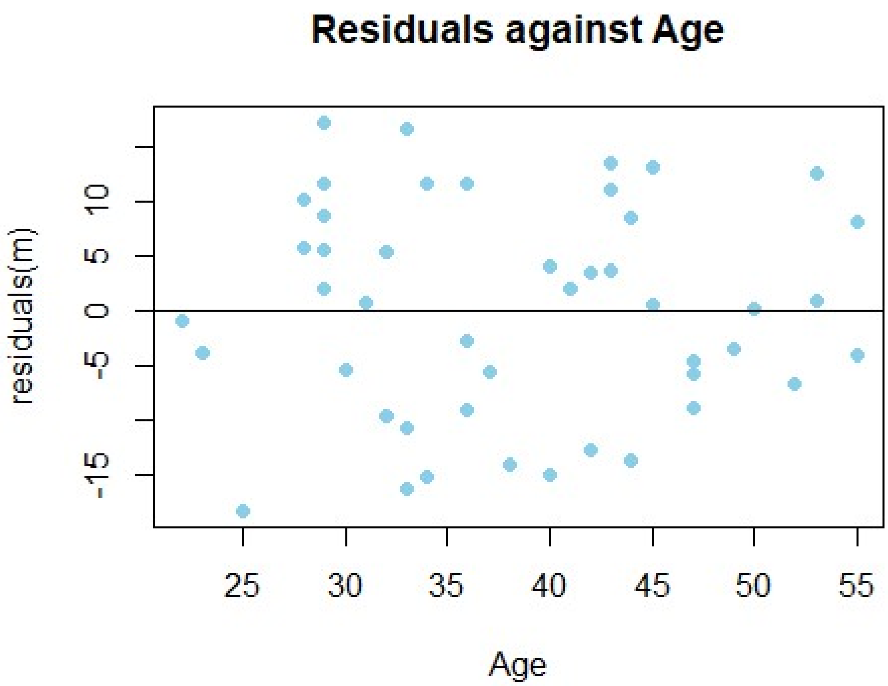
State the estimated regression function. How is b2 interpreted here

```
############### Fit regression model ###############
m <- lm(Satisfaction ~ Age + Illness + Anxiety, data=Satisfaction)
summary(m)

##
## Call:
## lm(formula = Satisfaction ~ Age + Illness + Anxiety, data = Satisfaction)
##
## 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 ***
## Age          -1.1416     0.2148  -5.315 3.81e-06 ***
## Illness      -0.4420     0.4920  -0.898   0.3741
## Anxiety     -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

# Satisfaction = β0 + β1*Age + β2*Illness + β3*Anxiety
# Satisfaction = 158.4913 -1.1416*Age -0.4420*Illness -13.4702*Anxiety

# Coefficient:
# The intercept (β0) is 158.4913, the estimated satisfaction when all predict
or variables are 0.

# For every one-year increase in age, the satisfaction decreases by 1.1416 un
its,
#     holding other variables constant.
# For every one-unit increase in the severity of illness, the satisfaction
#     decreases by 0.4420 units, holding other variables constant.
# For every one-unit increase in anxiety, the satisfaction decreases by
#     13.4702 units, holding other variables constant.

# With α=0.1, Anxiety(p=0.0647) and Age(p=3.81e-06) is significant, but not i
llness.
# Age & Anxiety are significant predictors for patient satisfaction.
# Anxiety has the strongest negative relationship with satisfaction among the
3 predictors.
```

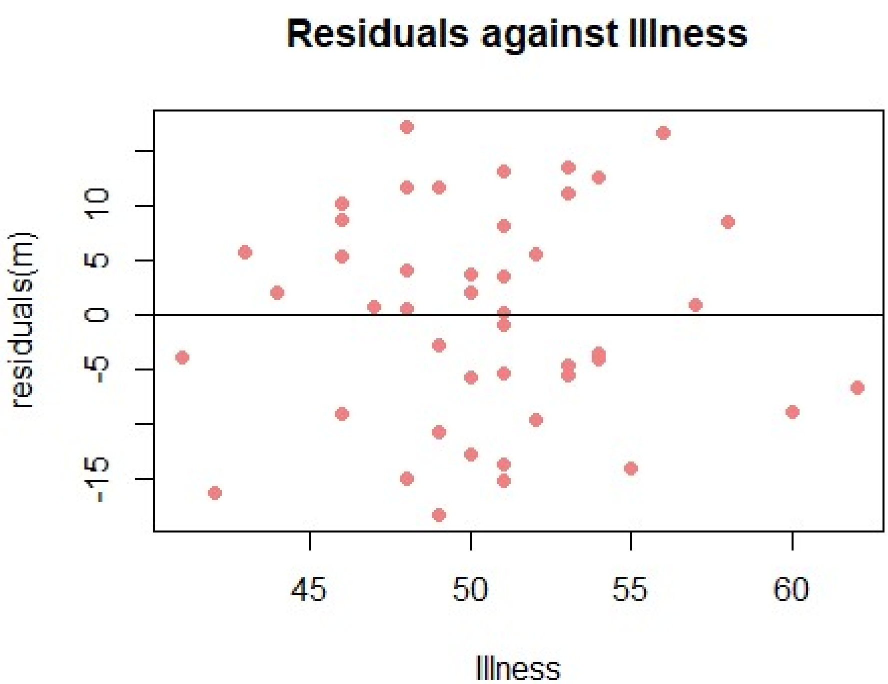
```
# Model-fit:
# 68.22% (R^2) of the variance in satisfaction is explained by the model.
# The multiple correlation coefficient r for the regression model is 0.825954
(sqrt 0.6822),
#       indicating a reasonably strong correlation between the dependent vari
able
#       (Sarisfaction) and the combined predictor variables (Age, illness, an
ziety).
# The model is statistically significant with p=1.542e-10 for the F-statistic
(30.05).

# Residual Analysis:
# The range from -18.3524 to 17.1601, and their median is close to zero (0.51
96).
# This suggests that the model doesn't systematically overestimate or
#                          underestimate satisfaction across the data.
```

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 iid~ N(0,σ^2)) is reasonable.

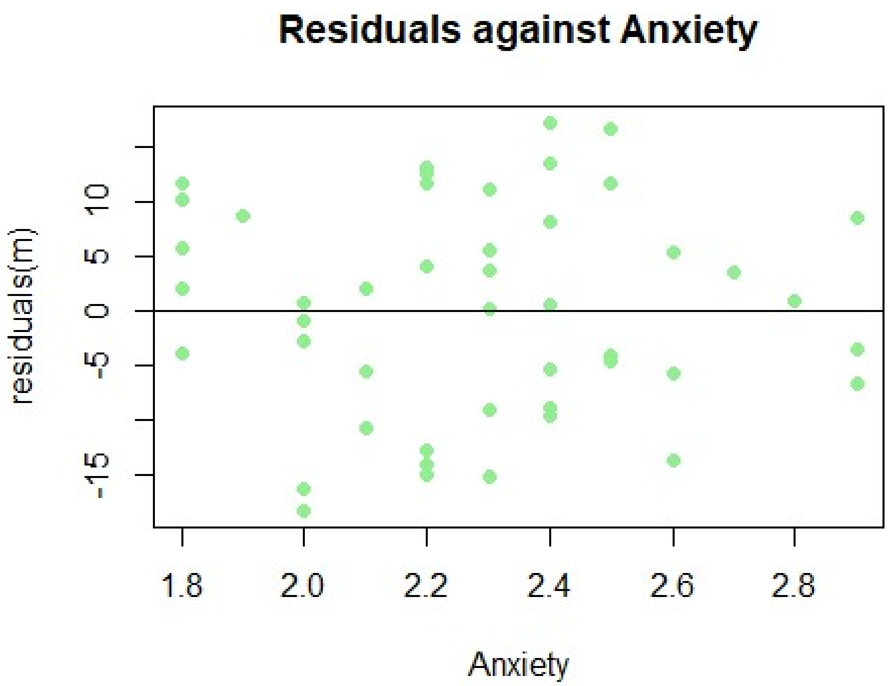
```
############### Residual diagnostics ###############

# (1) Check whether the relationship between Y and each X is linear.
# Plot the residuals against each X.
plot(Age, residuals(m), main = "Residuals against Age", pch=19, col="skyblue"
)
abline(h=0)
```



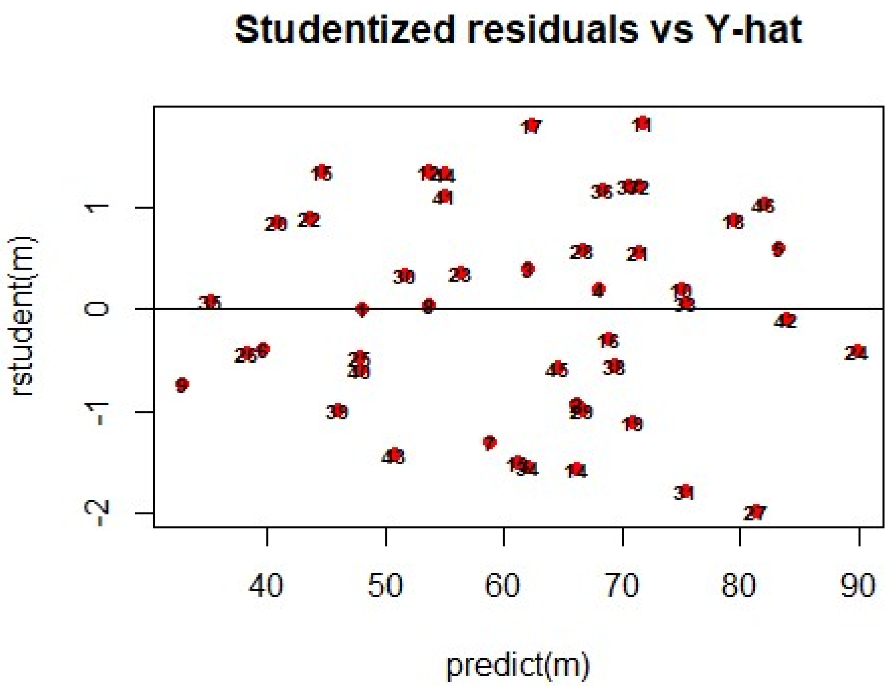

```
plot(Illness, residuals(m), main = "Residuals against Illness", pch=19, col="
lightcoral")
abline(h=0)
```



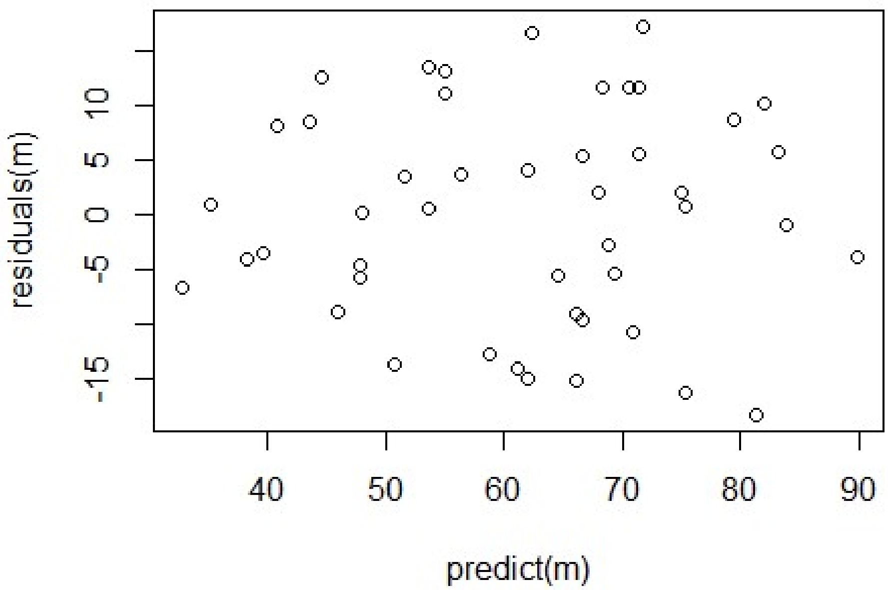
```
plot(Anxiety, residuals(m), main = "Residuals against Anxiety", pch=19, col="
lightgreen")
abline(h=0)
```



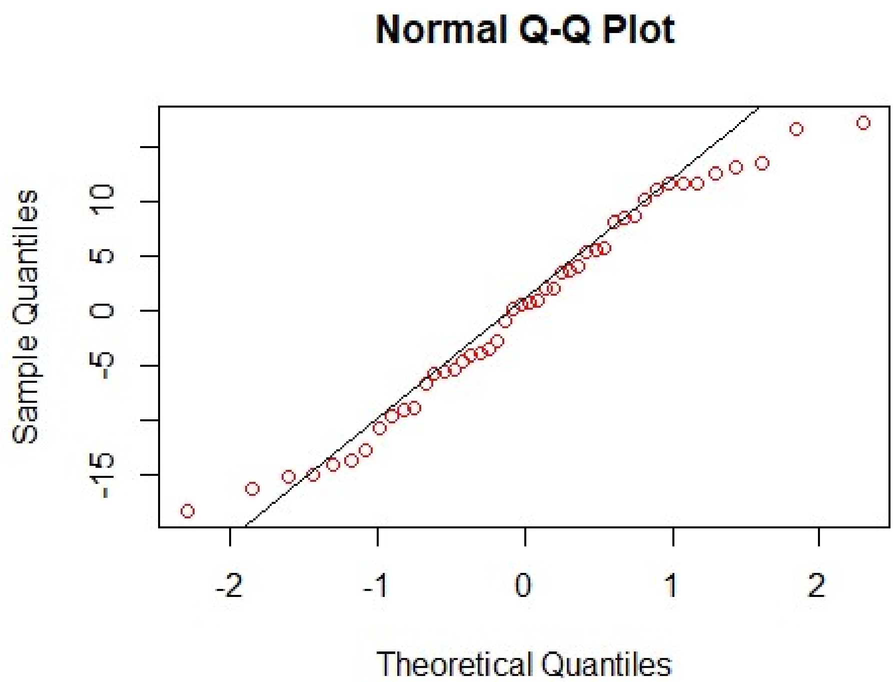

```
# Impression: The residuals do not seem to relate to any Age, Illness,
# or Anxiety in a systematic manner.
# Thus, the first-order terms of the 3 predictors in model m seems sufficient
.


# (2) Check for outliers.
# Plot the studentized residuals against Y-hat.
plot(predict(m), rstudent(m), main = "Studentized residuals vs Y-hat",
     pch=16, col="red")
text(predict(m), rstudent(m), names(rstudent(m)), cex=0.6, font=2)
abline(h=0)
```



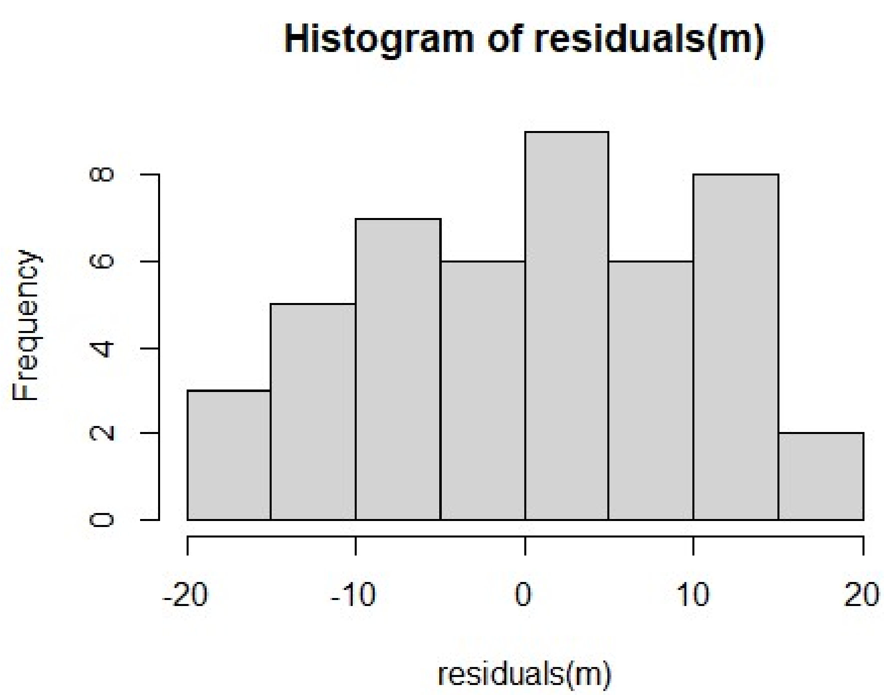

```
outliers <- rstudent(m)[abs(rstudent(m)) > 2] #potential outlier grater than
2
outliers

## named numeric(0)

# Impression: All studentized residuals are in a reasonable range given the
# approximate 1-2-3 rule.


# (3) Check for heteroskedasticity
# Plot the residuals against Y-hat
plot(predict(m), residuals(m))
```



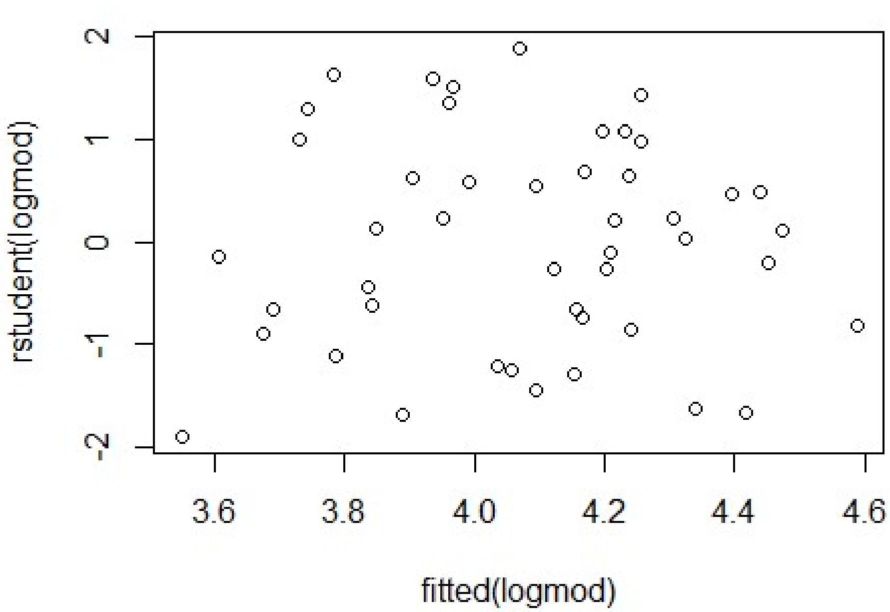

```r
# Impression: The vertical spread of the points are roughly constant across
# different X values. No concern of keteroskedasticity.


# (4) Check whether the residuals are normally distributed.
# Create the QQ-plot of the residuals.
qqnorm(residuals(m), col="red")
qqline(residuals(m)) # Our plot shows a slight deviation in the tails but
```



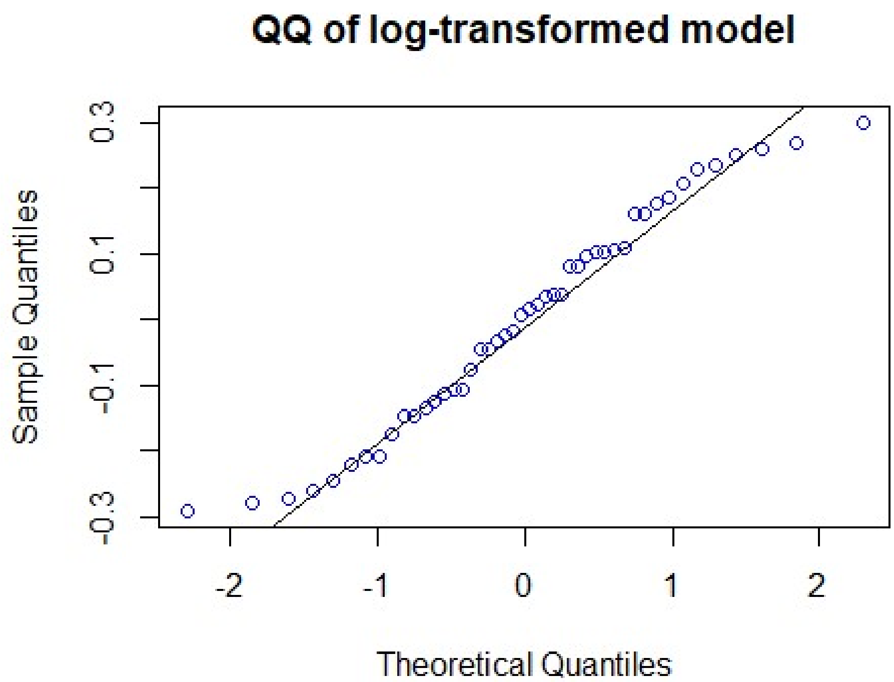

```
#                       is mostly aligned with the line, suggesting
#                       the residuals are approximately normally distributed.
hist(residuals(m), breaks=10)
```



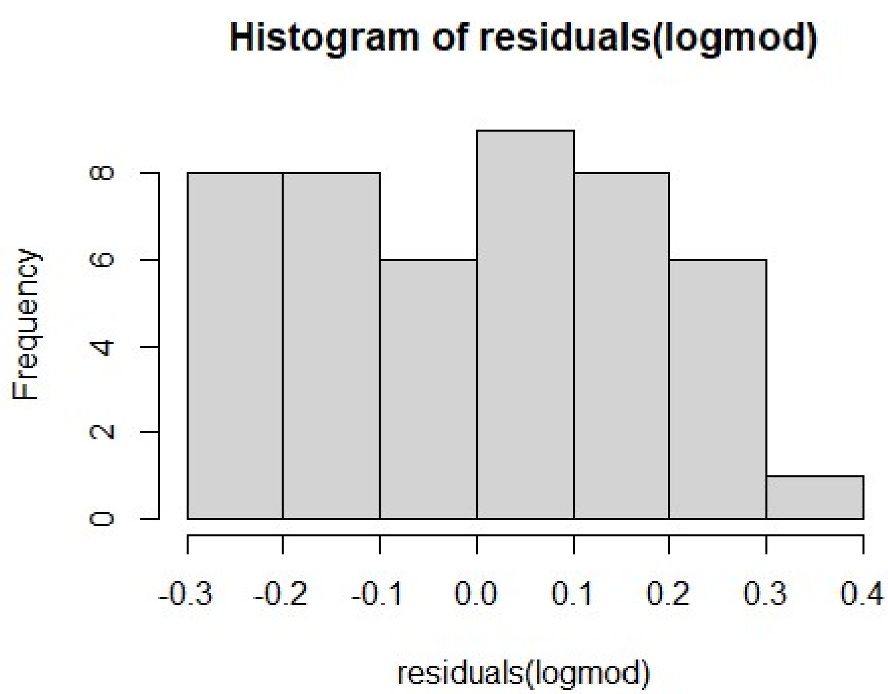
```
# log-transformed model
Satisfaction$log.Satisfaction <- log(Satisfaction$Satisfaction)
logmod <- lm(log.Satisfaction ~ Age + Illness + Anxiety, data=Satisfaction, n
a.action=na.exclude)
plot(fitted(logmod), rstudent(logmod)) # Residual plot of the log-transformed
model
```



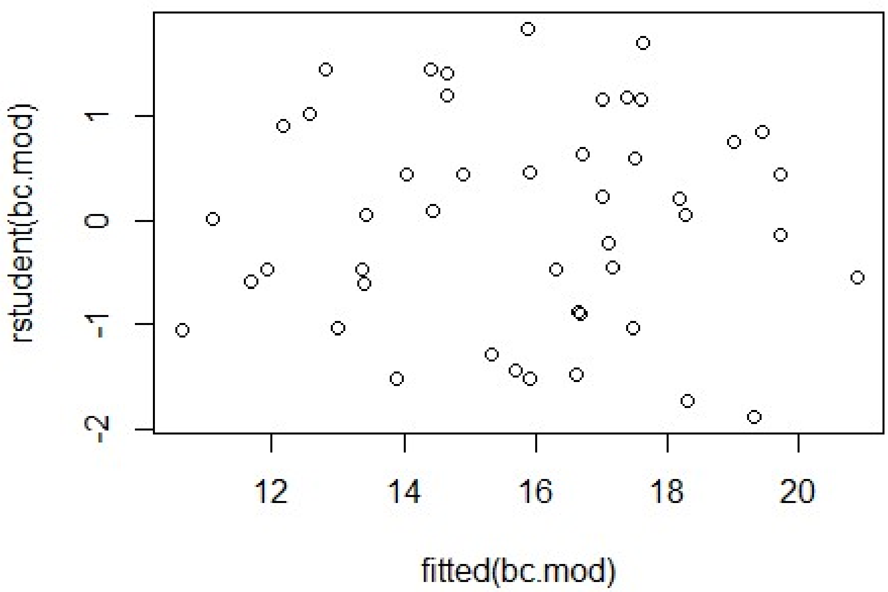

```
qqnorm(residuals(logmod), main="QQ of log-transformed model", col="blue") # Q
-Q plot of the log-transformed model
qqline(residuals(logmod)) # Residuals appear to deviate from the line, tail a
nd head
```



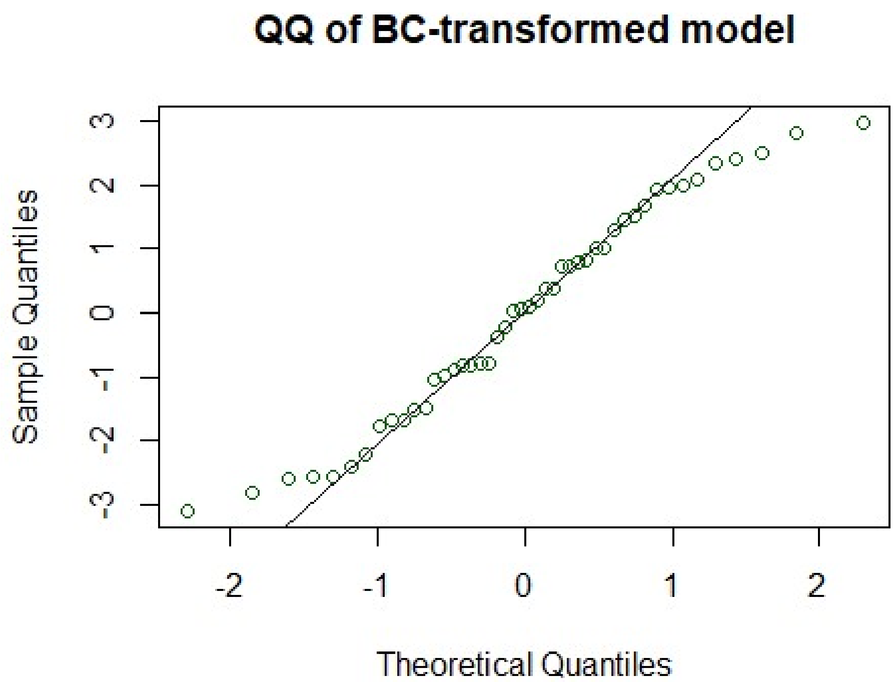

```
hist(residuals(logmod)) # skewed? Not bell-shape
```



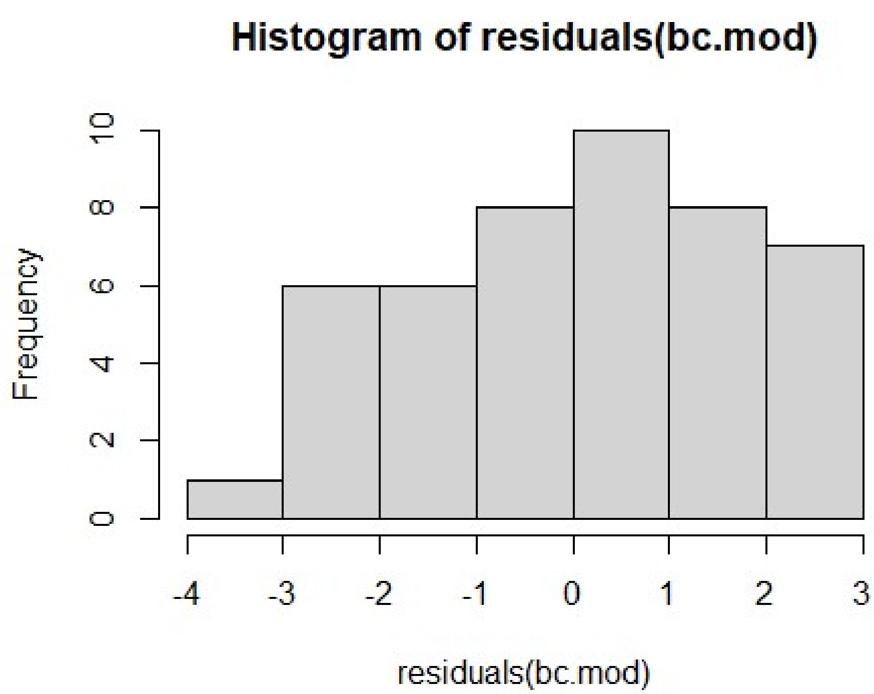
```
# Box-cox transformation: Transform Y with the lambda power and fit the model
again
lambda <- powerTransform(m)$lambda
lambda

##       Y1
## 0.672265

Satisfaction$bc.Satisfaction <- Satisfaction$Satisfaction ^ lambda
bc.mod <- lm(bc.Satisfaction ~ Age + Illness + Anxiety, data=Satisfaction, na
.action=na.exclude)
plot(fitted(bc.mod), rstudent(bc.mod)) # Residual plot of the box-cox-transfo
rmed model
```




```
qqnorm(residuals(bc.mod), main="QQ of BC-transformed model", col="darkgreen")
# Q-Q plot of the box-cox-transformed model
qqline(residuals(bc.mod)) # Residuals appear to deviate from the line, tail a
nd head
```




```
hist(residuals(bc.mod)) # skewed. Distribution is not well bell-curved.
```



```
# Shapiro test.
# ncvTest() function from the car package performs a non-constant variance sc
ore test
# (known as the Breusch-Pagan test) to test for heteroscedasticity in a linea
r regression model.
# H0: the variances of the residuals are constant (homoscedasticity),
# H1: they are not constant (heteroscedasticity).
ncvTest(logmod)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.9292462, Df = 1, p = 0.33506

ncvTest(bc.mod)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.08271333, Df = 1, p = 0.77365

ncvTest(m)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6513299, Df = 1, p = 0.41964

# impression: p=value for all 3 transformed models are high.
# There's no evidence to reject H0, meaning homoscedastic.


# (5) whether the observations are independent from each other
# Each row in the data represents a patient.
# Also, residuals fluctuate in a more or less random pattern around the basel
ine 0.
# Thus, it is reasonable to assume the patients are independent with respect
to
# the three measures involved in this regression problem.

# Summary:the assumptions of the normal error regression (NER) model
# (i.e., ε iid~ N(0,σ^2)) is reasonable.
# (1) The relationship between Y and each X is linear,
# (2) No outlier
# (3) Homoscedastic: The variance of the errors is constant across all levels
of the independent variables.
# (4) ε (0,σ^2): slightly deviated in the tails but is mostly aligned with th
e line,
#              suggesting the residuals are approximately normally distributed
.
# (5) independent observations: patients are independent
```

CI for coefficient

```r
############### CI for coefficient ###############
# b1 +/- qt*s(bi)
qt(p=1-0.1/2, 46-4)

## [1] 1.681952

# 90% CI for β1, β2, β3
confint(m, level = 0.9)

##                      5 %        95 %
## (Intercept) 128.004370 188.9781330
## Age          -1.502893  -0.7803305
## Illness      -1.269467   0.3854587
## Anxiety     -25.411454  -1.5288719
```