A3

Jefferson Li, Arib Shaikh

09/03/2021

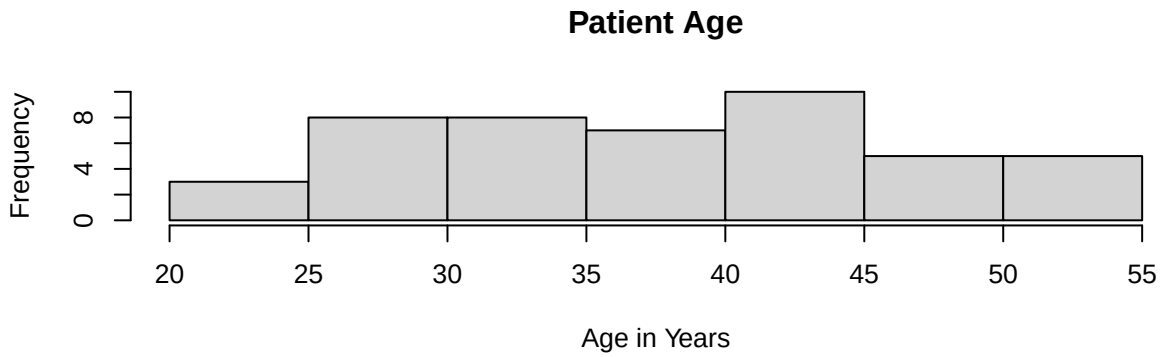
Q1

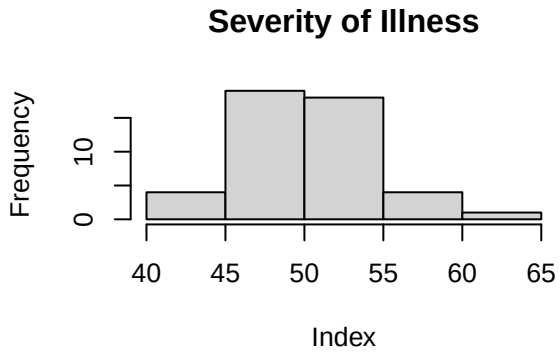
a)

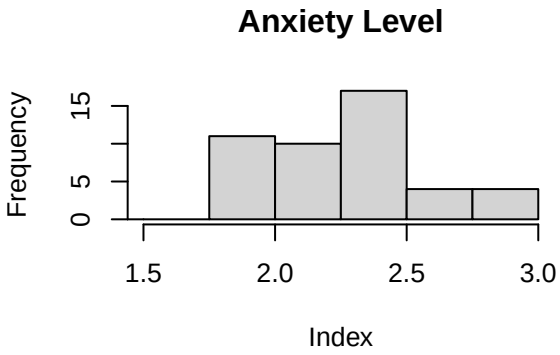
```
Data = read.table("PatientSatisfaction.txt",  col.names=c("Y", "X1", "X2", "X3"))

layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(Data$X1,
  main="Patient Age",
  xlab="Age in Years",
  xlim=c(20,55),
)
hist(Data$X2,
  main="Severity of Illness",
  xlab="Index",
  xlim=c(40,65),
)
hist(Data$X3,
  main="Anxiety Level",
  xlab="Index",
  xlim=c(1.5,3),
  breaks=c(1,1.25,1.5,1.75,2,2.25,2.5,2.75,3)

)
```





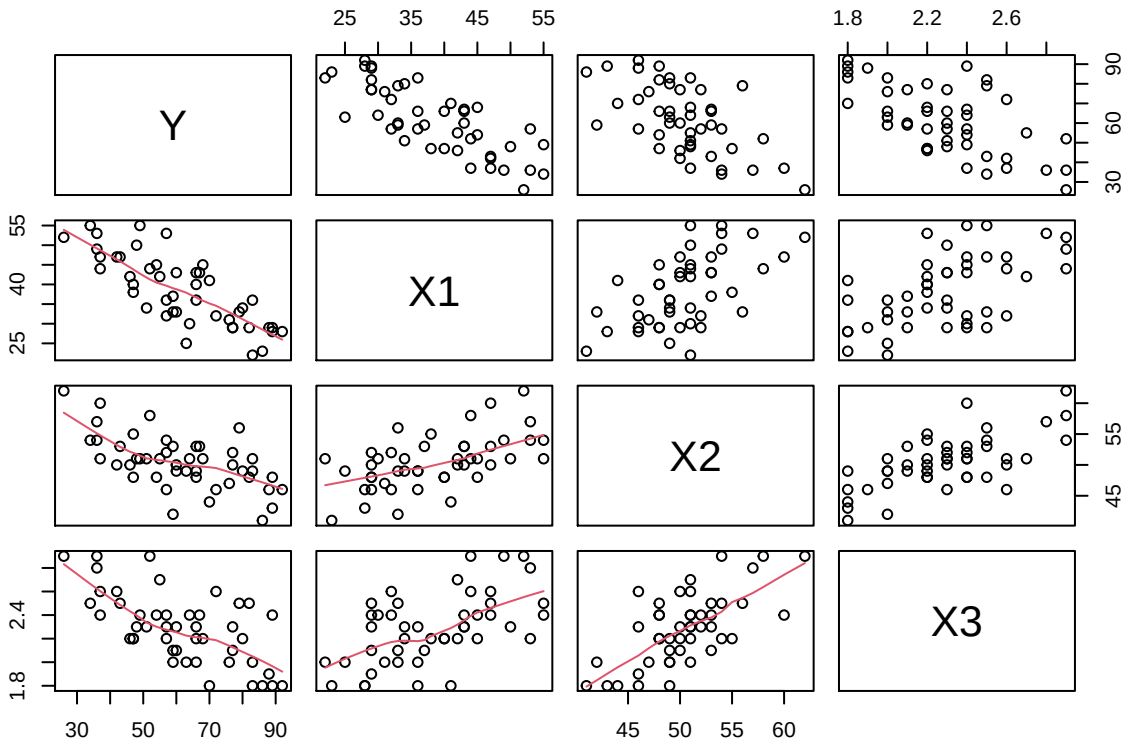


It is noteworthy that it seems all 3 plots are normally distributed.

b)

```
# scatter plot matrix

pairs(~Y +X1 + X2+ X3, data = Data, lower.panel = panel.smooth)
```




```
cor(cbind(Data$X1, Data$X2, Data$X3))
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5679505 0.5696775
## [2,] 0.5679505 1.0000000 0.6705287
## [3,] 0.5696775 0.6705287 1.0000000
```

Our scatter plot matrix shows that all 3 predictor value are positively correlated with each other, however, all 3 are negatively correlated with patient satisfaction.

Since none of the correlations between the predictor variables exceed 0.7, the correlations are not extreme enough to raise any concerns of multicollinearity.

c)

```
fit = lm(Y ~ X1 + X2 + X3, data = Data)
summary(fit)
```

```
##
## 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

$$Y = 158.4912517 + -1.1416118 \cdot X_1 + -0.4420043 \cdot X_2 + -13.4701632 \cdot X_3$$

$\hat{\beta}_2$ is interpreted as, controlling for X_1 and X_3, a 1-unit increase in X_2 corresponds to a predicted increase of -1.1416118 in patient satisfaction.

d)

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0, \quad H_a : \hat{\beta}_1 \neq 0 \vee \hat{\beta}_2 \neq 0 \vee \hat{\beta}_3 \neq 0$$

Decision rule : reject if

$$\text{P Value} = pf(F, 3, n - p') < \alpha$$

then we reject H_0

```
alpha = 0.10

anova = anova(fit)
anova
```

```
## 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
```

```
SSR =  sum(anova[2][1:3,1])
MSE = anova[3][4,1]

Fstat = SSR/MSE/3
Fstat
```

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

$$Fstatistic = 30.0520779$$

```
Pval = pf(Fstat,4-1, anova[1][4,1],lower.tail=F)
Pval
```

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

Since the P value of $1.5419726 \times 10^{-10}$ is less than the level of significance of $\alpha = 0.1$we can reject H_0

My test implies that we are 90% confident that $\beta_1, \beta_2, \beta_3$ are not all 0.

e)

```
SSE = anova[2][4,1]
SST = (SSR + SSE)
SSE
```

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

```
R = SSR/SST
R
```

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

```
Radj = 1 - ((anova[1][4,1]+4-1)/anova[1][4,1])*(SSE/SST)
Radj
```

```
## [1] 0.6594939
```

The R^2 value of 0.6821943 means that $100 \cdot (0.6821943)\%$ of the data fit the regression model. Furthermore, since the adjusted R^2 value is close to R^2, we can say most prediction variables are having an effect on predictions.

f)

```
#fit
pred = predict(fit, data.frame(X1 = 35, X2 = 45, X3 = 2.2 ) , interval="prediction")
pred
```

```
##        fit      lwr      upr
## 1 69.01029 48.01224 90.00833
```

I am 95% confident that a future patient with age 35, severity of illness 45, and anxiety level 2.2 will have a patient satisfaction within the range

$$[48.0122427, 90.0083297]$$