STAT 5385: Lab 4

Willliam Ofosu Agyapong

24/02/2022

1 Introducing the SENIC data

Table 1: Variables in the SENIC dataset

Variable Name	Coded As
Identification Number	ID
Length of stay	LOS
Age	age
Infection risk	infec_risk
Routine culturing ratio	cul_ratio
Routine chest X-ray ratio	xray_ratio
Number of beds	beds
Medical school affiliation	med
Region	region
Average daily census	ADC
Number of nurses	nurses
Available facilities and services	AFS

1.1 Reading in the dataset

```
# Import the dataset from local drive
senic <- read.table(file = "../Data Sets/Appendix C Data Sets/APPENC01.txt",
                    header = FALSE)

# Rename columns
names(senic) <- var_table$coded_names

# View first few observations; Everything looks good.
head(senic)
```

```
##   ID   LOS  age infec_risk cul_ratio xray_ratio beds med region ADC nurses AFS
## 1  1  7.13 55.7        4.1       9.0       39.6  279   2      4 207    241  60
## 2  2  8.82 58.2        1.6       3.8       51.7   80   2      2  51     52  40
## 3  3  8.34 56.9        2.7       8.1       74.0  107   2      3  82     54  20
## 4  4  8.95 53.7        5.6      18.9      122.8  147   2      4  53    148  40
## 5  5 11.20 56.5        5.7      34.5       88.9  180   2      1 134    151  40
## 6  6  9.76 50.9        5.1      21.9       97.0  150   2      2 147    106  40
```

```
# Obtain summary report
# senic %>% dfSummary() %>% view()
```

2 Problem 3.27

2.1 Part (a): Diagnostic Plots

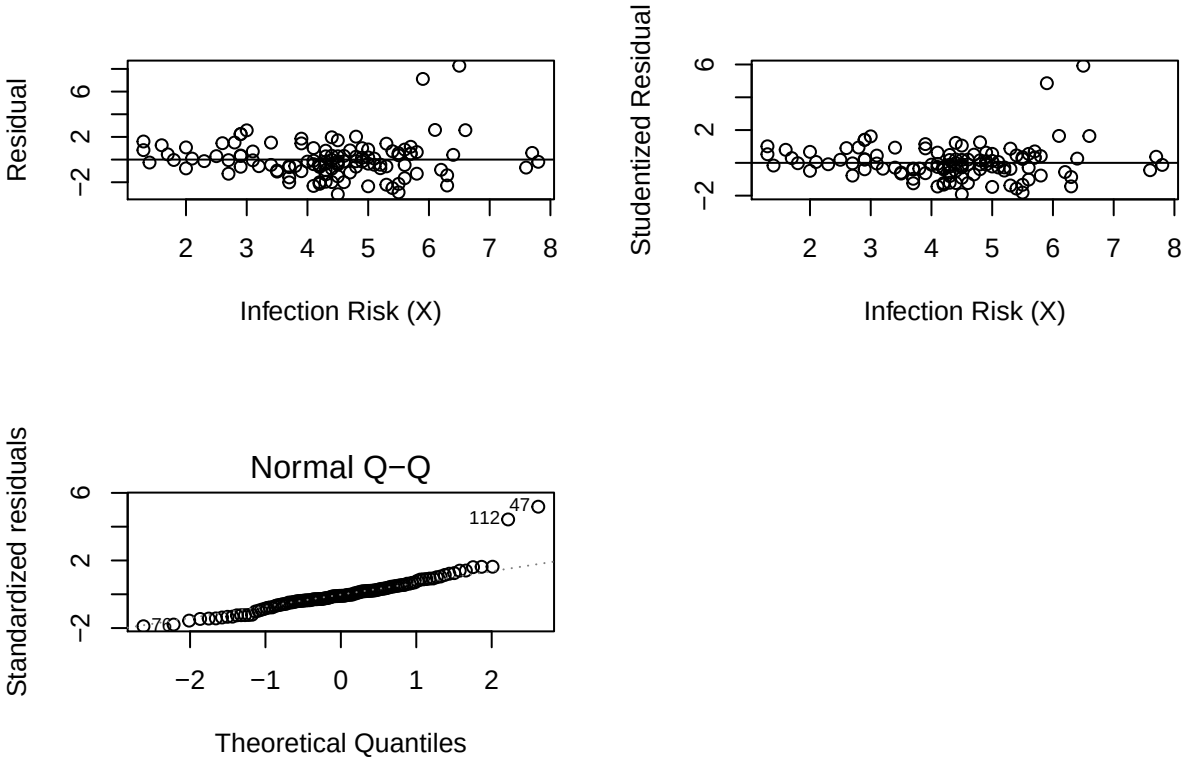
2.1.1 Infection Risk

```
# Fit a linear model
mod_infec_risk <- lm(LOS ~ infec_risk, data = senic)

# residuals against infection risk
 par(mfrow=c(2,2))

plot(senic$infec_risk, resid(mod_infec_risk),
     xlab = "Infection Risk (X)", ylab = "Residual"); abline(h=0)

plot(senic$infec_risk, rstudent(mod_infec_risk),
     xlab = "Infection Risk (X)", ylab = "Studentized Residual"); abline(h=0)

# normal probability plot
plot(mod_infec_risk, which = 2)

# investigating outliers
summary(senic[, c("LOS", "infec_risk")])
```

```
##       LOS          infec_risk
##  Min.   : 6.70   Min.   :1.30
##  1st Qu.: 8.34   1st Qu.:3.70
##  Median : 9.42   Median :4.40
##  Mean   : 9.65   Mean   :4.36
##  3rd Qu.:10.47   3rd Qu.:5.20
##  Max.   :19.56   Max.   :7.80
```

```
# look at potential outliers: 47 and 112
senic[c(47,76,112), c("LOS", "infec_risk")]
```

```
##       LOS infec_risk
## 47  19.56        6.5
## 76   6.70        4.5
## 112 17.94        5.9
```



- Potential outliers, Looking at the residual versus the majority of observations are between 4 and 6 rate of infection risk. They can influence the kind of transformation
-

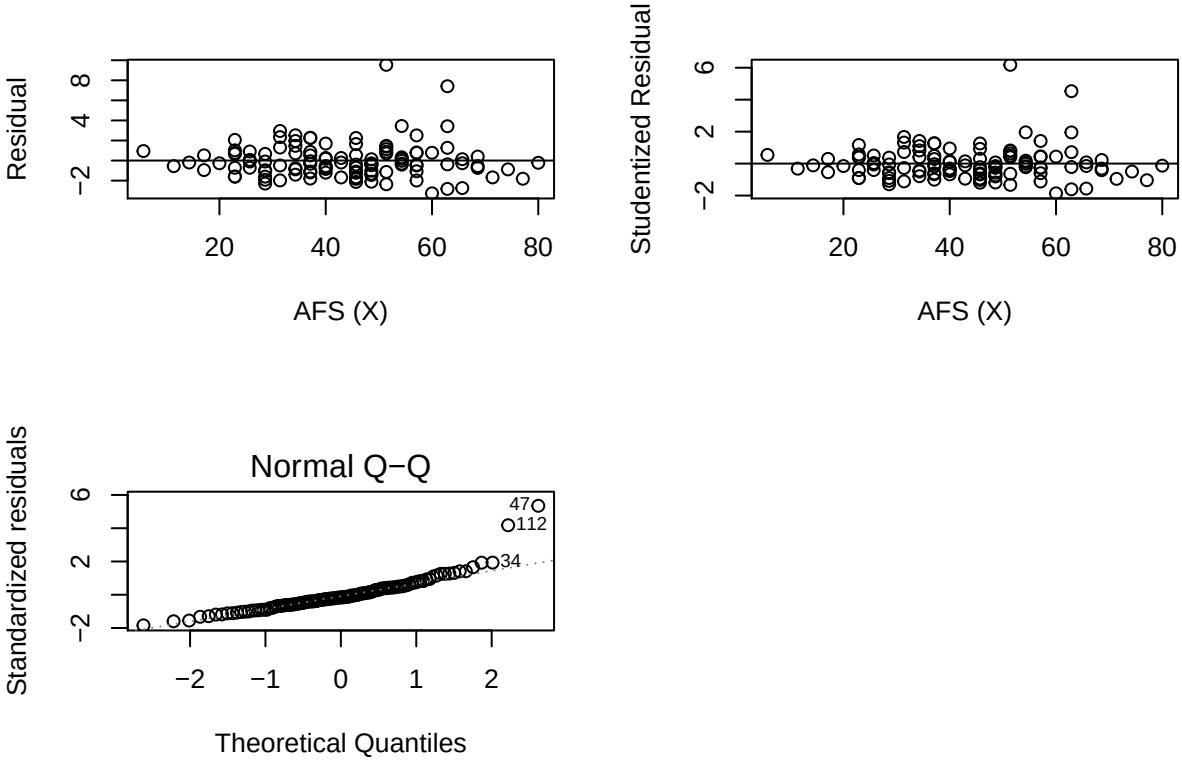
2.1.2 Available Facilities and Services (AFS)

```
mod_AFS <- lm(LOS ~ AFS, data = senic)
# residuals agains infection risk
par(mfrow = c(2,2))
plot(senic$AFS, resid(mod_AFS),
     xlab = "AFS (X)", ylab = "Residual"); abline(h=0)

plot(senic$AFS, rstudent(mod_AFS),
     xlab = "AFS (X)", ylab = "Studentized Residual"); abline(h=0)

# normal probability plot
plot(mod_AFS, which = 2)

# investigating outliers
summary(senic[, c("LOS", "AFS")])
```

```
##       LOS             AFS
##  Min.   : 6.70   Min.   : 5.7
##  1st Qu.: 8.34   1st Qu.:31.4
##  Median : 9.42   Median :42.9
##  Mean   : 9.65   Mean   :43.2
##  3rd Qu.:10.47   3rd Qu.:54.3
##  Max.   :19.56   Max.   :80.0
```

```
# look at potential outliers: 47 and 112
senic[c(34,47,112), c("LOS", "AFS")]
```

```
##       LOS  AFS
## 34  13.59 54.3
## 47  19.56 51.4
## 112 17.94 62.9
```



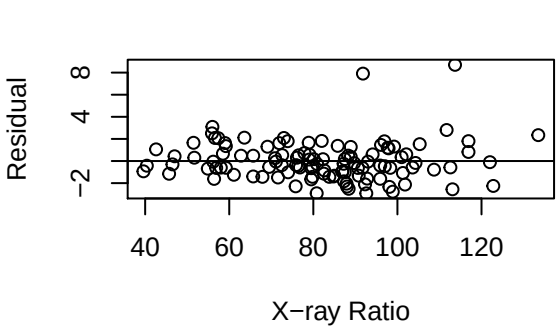
2.1.3 Chest X-ray Ratio

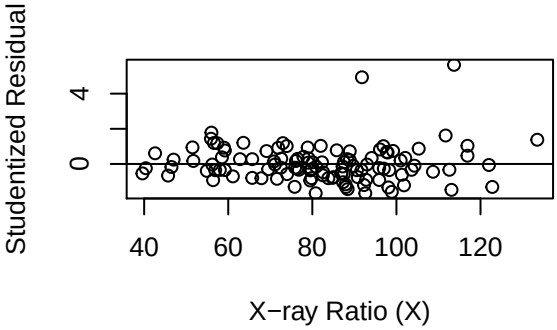
```
mod_xray_ratio <- lm(LOS ~ xray_ratio, data = senic)
# residuals agains infection risk
par(mfrow = c(2,2))
plot(senic$xray_ratio, resid(mod_xray_ratio),
     xlab = "X-ray Ratio", ylab = "Residual"); abline(h=0)

plot(senic$xray_ratio, rstudent(mod_xray_ratio),
     xlab = "X-ray Ratio (X)", ylab = "Studentized Residual"); abline(h=0)

# normal probability plot
plot(mod_xray_ratio, which = 2)

# investigating outliers
summary(senic[, c("LOS", "xray_ratio")])
```

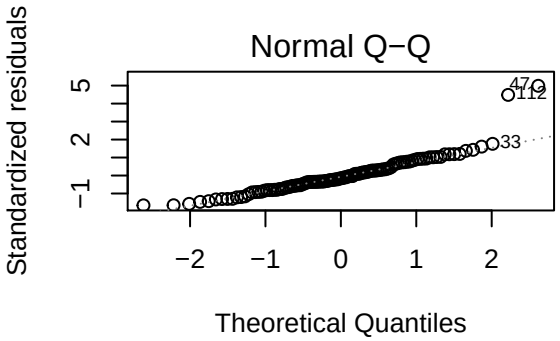
```
##       LOS          xray_ratio
##  Min.   : 6.70   Min.   : 39.6
##  1st Qu.: 8.34   1st Qu.: 69.5
##  Median : 9.42   Median : 82.3
##  Mean   : 9.65   Mean   : 81.6
##  3rd Qu.:10.47   3rd Qu.: 94.1
##  Max.   :19.56   Max.   :133.5
```

```
# look at potential outliers: 47 and 112
senic[c(33,47,112), c("LOS", "xray_ratio")]
```

```
##       LOS xray_ratio
## 33  11.77       56.0
## 47  19.56      113.7
## 112 17.94       91.8
```







- Two data points at the 47th and 112th observations were identified as obvious outliers in all three models.
- The two extreme observations appear to be the main factor contributing to the of lack of fit in all the models.
- Therefore, except for the outliers causing some departures, a linear regression appears to be slightly more appropriate in the cases involving X-ray ratio and AFS as predictors than the model involving infection risk.

2.2 Part (b): Models without outlying observations

Observations **47** and **112** appeared to be extreme values for all the three models involving the three individual predictors. As a next step, we refit all models without these two observations to assess their impact on model performance.

2.2.1 Infection Risk

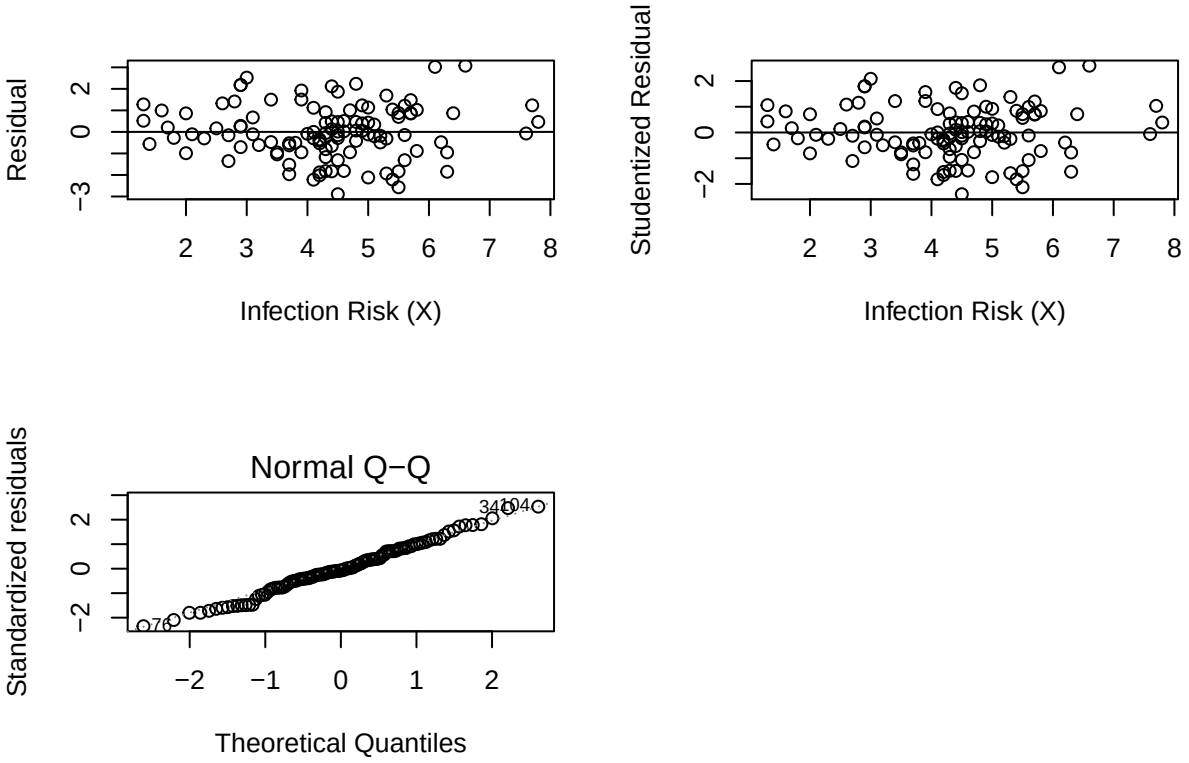
```
senic2 <- senic[-c(47, 112),] # data without outliers
mod_infec_risk_b <- lm(LOS ~ infec_risk, data = senic2)
# summary(mod_infec_risk_b)

# residuals agains infection risk
par(mfrow = c(2,2))
plot(senic2$infec_risk, resid(mod_infec_risk_b),
     xlab ="Infection Risk (X)",ylab ="Residual"); abline(h=0)

plot(senic2$infec_risk, rstudent(mod_infec_risk_b),
     xlab ="Infection Risk (X)",ylab ="Studentized Residual"); abline(h=0)

# normal probability plot
plot(mod_infec_risk_b, which = 2)
```

```
# investigating outliers
# look at
summary(senic[, c("LOS","infec_risk")])
```

```
##       LOS          infec_risk
##  Min.   : 6.70   Min.   :1.30
##  1st Qu.: 8.34   1st Qu.:3.70
##  Median : 9.42   Median :4.40
##  Mean   : 9.65   Mean   :4.36
##  3rd Qu.:10.47   3rd Qu.:5.20
##  Max.   :19.56   Max.   :7.80
```

```
senic2[c(34,104), c("LOS","infec_risk")]
```

```
##       LOS infec_risk
## 34  13.59        6.1
## 105  9.44        4.5
```




```
# obtaining 95% prediction interval at Xh = 6.5
kable(predict(mod_infec_risk_b, newdata = data.frame(infec_risk = 6.5), interval = "prediction"))
```

2.2.1.1 95% Prediction Intervals

fit	lwr	upr
10.81	8.319	13.31

```
# obtaining 95% prediction interval at Xh = 5.9
kable(predict(mod_infec_risk_b, newdata = data.frame(infec_risk = 5.9), interval = "prediction"))
```

fit	lwr	upr
10.45	7.967	12.93

Clearly, observations $Y_{47} = 19.56$ and $Y_{112} = 17.94$ lie far outside the two prediction intervals corresponding to $X_h = 6.5$ and $X_h = 5.9$, respectively.

The fact that the observations fall outside the prediction intervals could be an indication that those observations most likely arose from measurement errors.

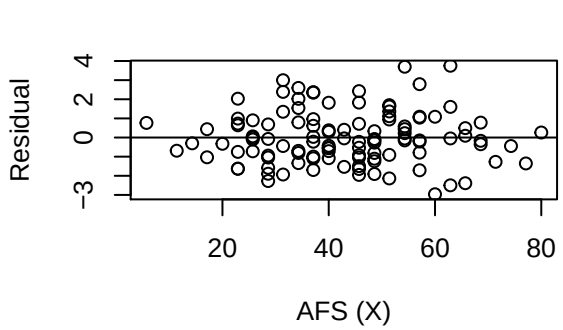
2.2.2 Available Facilities and Services (AFS)

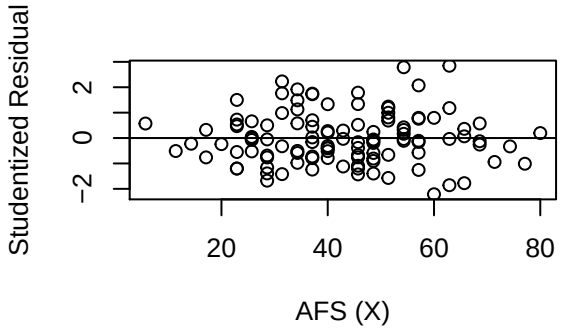
```
senic2 <- senic[-c(47, 112),] # data without outliers
mod_AFS_b <- lm(LOS ~ AFS, data = senic2)
# summary(mod_AFS_b)

# residuals agains infection risk
par(mfrow = c(2,2))
plot(senic2$AFS, resid(mod_AFS_b),
     xlab ="AFS (X)",ylab ="Residual"); abline(h=0)

plot(senic2$AFS, rstudent(mod_AFS_b),
     xlab ="AFS (X)",ylab ="Studentized Residual"); abline(h=0)

# normal probability plot
plot(mod_AFS_b, which = 2)
# investigating outliers
# look at
summary(senic[, c("LOS","AFS")])
```

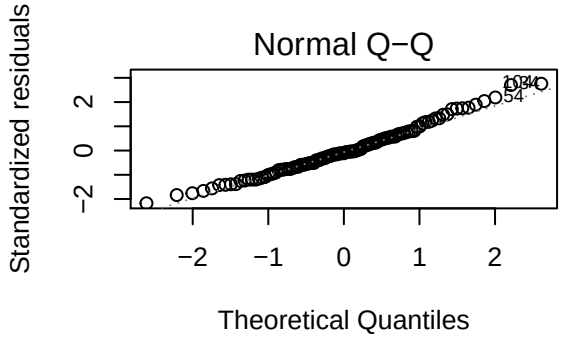
```
##       LOS             AFS
##  Min.   : 6.70   Min.   : 5.7
##  1st Qu.: 8.34   1st Qu.:31.4
##  Median : 9.42   Median :42.9
##  Mean   : 9.65   Mean   :43.2
##  3rd Qu.:10.47   3rd Qu.:54.3
##  Max.   :19.56   Max.   :80.0
```

```
senic2[c(34,104), c("LOS","AFS")]
```

```
##       LOS  AFS
## 34  13.59 54.3
## 105  9.44 42.9
```








```
# obtaining 95% prediction interval at Xh = 51.4
kable(predict(mod_AFS_b, newdata = data.frame(AFS = 51.4), interval = "prediction"))
```

2.2.2.1 95% Prediction Intervals

fit	lwr	upr
9.787	7.041	12.53

```
# obtaining 95% prediction interval at Xh = 62.9
kable(predict(mod_AFS_b, newdata = data.frame(AFS = 62.9), interval = "prediction"))
```

fit	lwr	upr
10.2	7.434	12.96

The same observation is made here. The observations fall outside the prediction intervals.

2.2.3 Chest X-ray Ratio

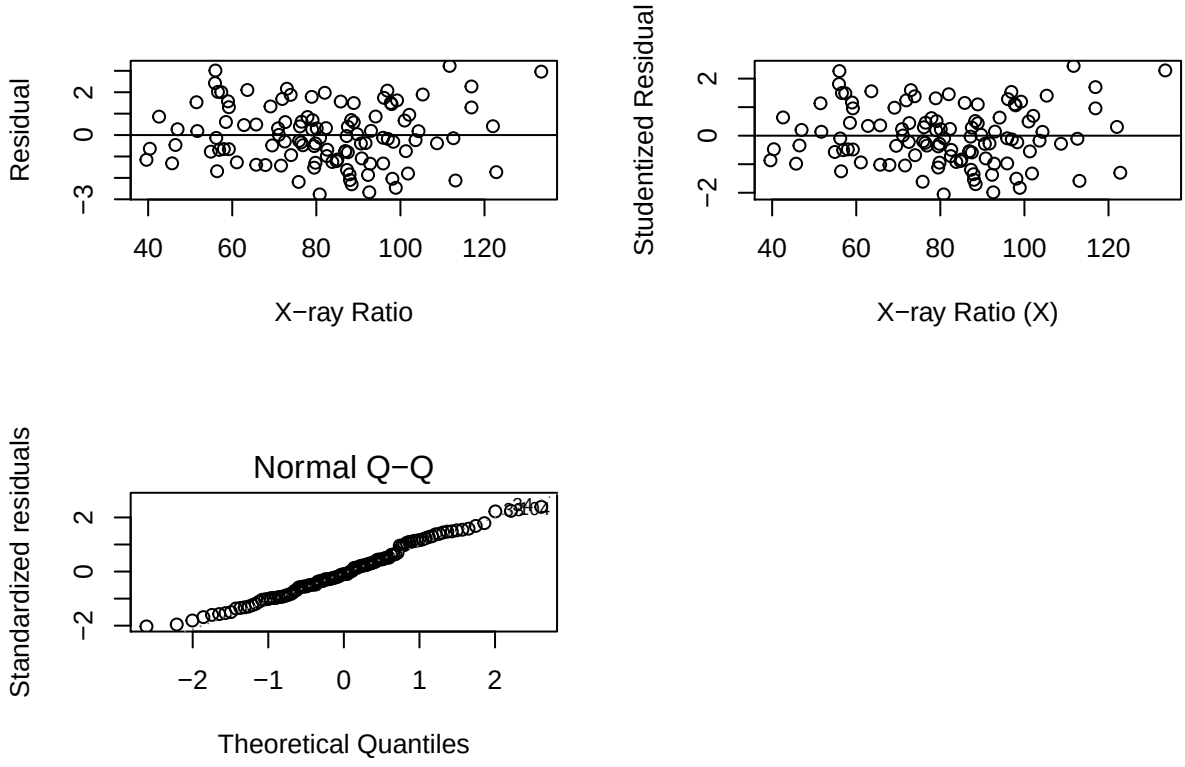
```
senic2 <- senic[-c(47, 112),] # data without outliers
mod_xray_ratio_b <- lm(LOS ~ xray_ratio, data = senic2)
# summary(mod_AFS_b)

# residuals agains infection risk
par(mfrow = c(2,2))
plot(senic2$xray_ratio, resid(mod_xray_ratio_b),
     xlab ="X-ray Ratio",ylab ="Residual"); abline(h=0)
```

```
plot(senic2$xray_ratio, rstudent(mod_xray_ratio_b),
     xlab ="X-ray Ratio (X)",ylab ="Studentized Residual"); abline(h=0)

# normal probability plot
plot(mod_xray_ratio_b, which = 2)
# investigating outliers
# look at
summary(senic2[, c("LOS","xray_ratio")])
```

```
##       LOS          xray_ratio
##  Min.   : 6.70   Min.   : 39.6
##  1st Qu.: 8.32   1st Qu.: 69.3
##  Median : 9.41   Median : 82.0
##  Mean   : 9.48   Mean   : 81.2
##  3rd Qu.:10.40   3rd Qu.: 93.5
##  Max.   :13.95   Max.   :133.5
```

```
senic2[c(34,104), c("LOS","xray_ratio")]
```

```
##       LOS xray_ratio
## 34  13.59      111.7
## 105  9.44       58.5
```




```
# obtaining 95% prediction interval at Xh = 113.7
kable(predict(mod_xray_ratio_b, newdata = data.frame(xray_ratio = 113.7), interval = "prediction"))
```

2.2.3.1 95% Prediction Intervals

fit	lwr	upr
10.42	7.652	13.19

```
# obtaining 95% prediction interval at Xh = 91.8
kable(predict(mod_xray_ratio_b, newdata = data.frame(xray_ratio = 91.8), interval = "prediction"))
```

fit	lwr	upr
9.788	7.052	12.52

As noted earlier in the other cases, the observations are also outside these prediction intervals, and this could be a signal that the observations are that extreme due to possible measurement errors

When comparing the models

3 Problem 3.28

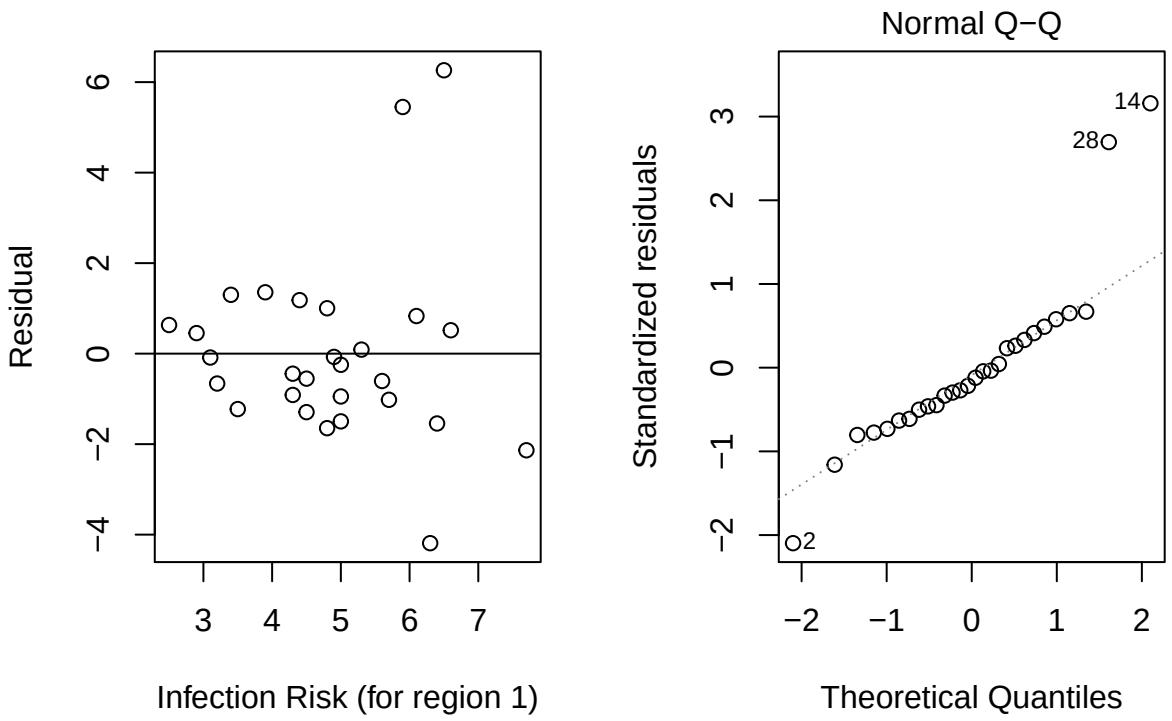
```
models <- list() # initialize empty list

# Get the various regions
regions <- as.integer(levels(as.factor(senic$region)))

# fit individual models for each region
for(i in regions) {
  models[[i]] = senic %>%
filter(region == i) %>%
    lm(LOS ~ infec_risk, data = .)
  # print(paste("Estimated model summary for region", i))
  # print(summary(models[[i]]))
}
```

3.1 Diagnostic plots

We look at diagnostic plots for models involving Length of Stay (Y) and Infection Risk for each region.

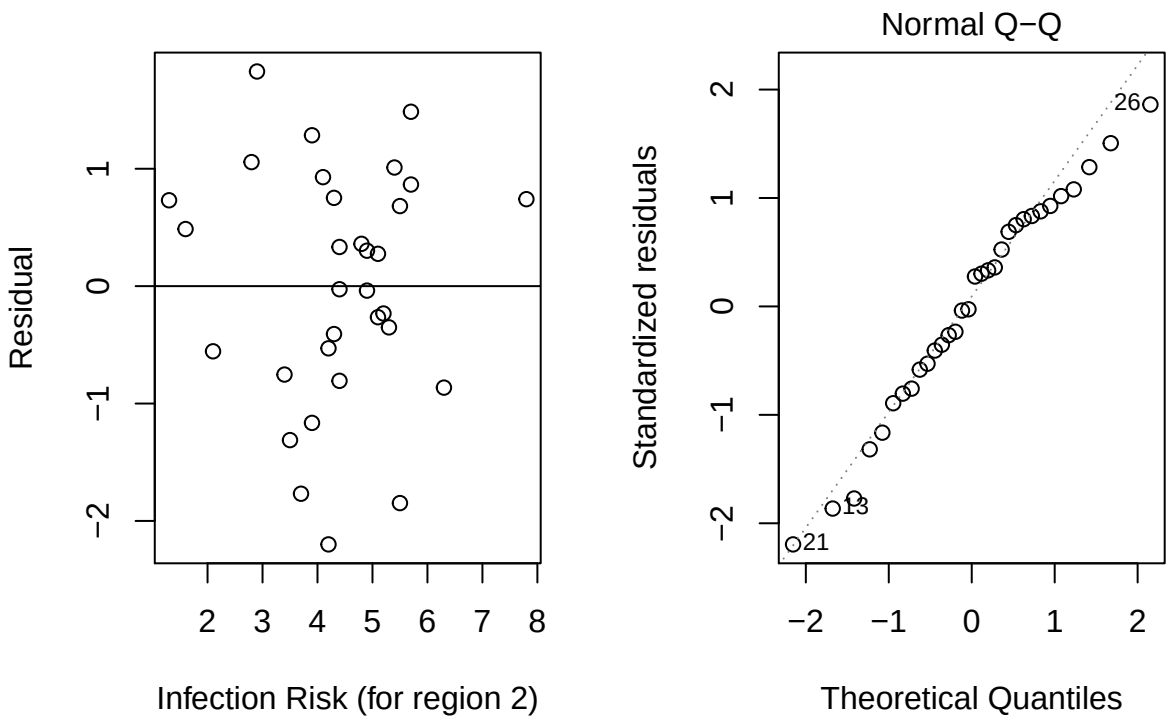
3.1.1 Region 1

```
# region 1
par(mfrow = c(1,2))
plot(senic[senic$region==1,]$infec_risk, resid(models[[1]]),
     xlab ="Infection Risk (for region 1)",ylab = "Residual"); abline(h=0) # residuals against infection r
plot(models[[1]], which = 2)# normal probability plot
```



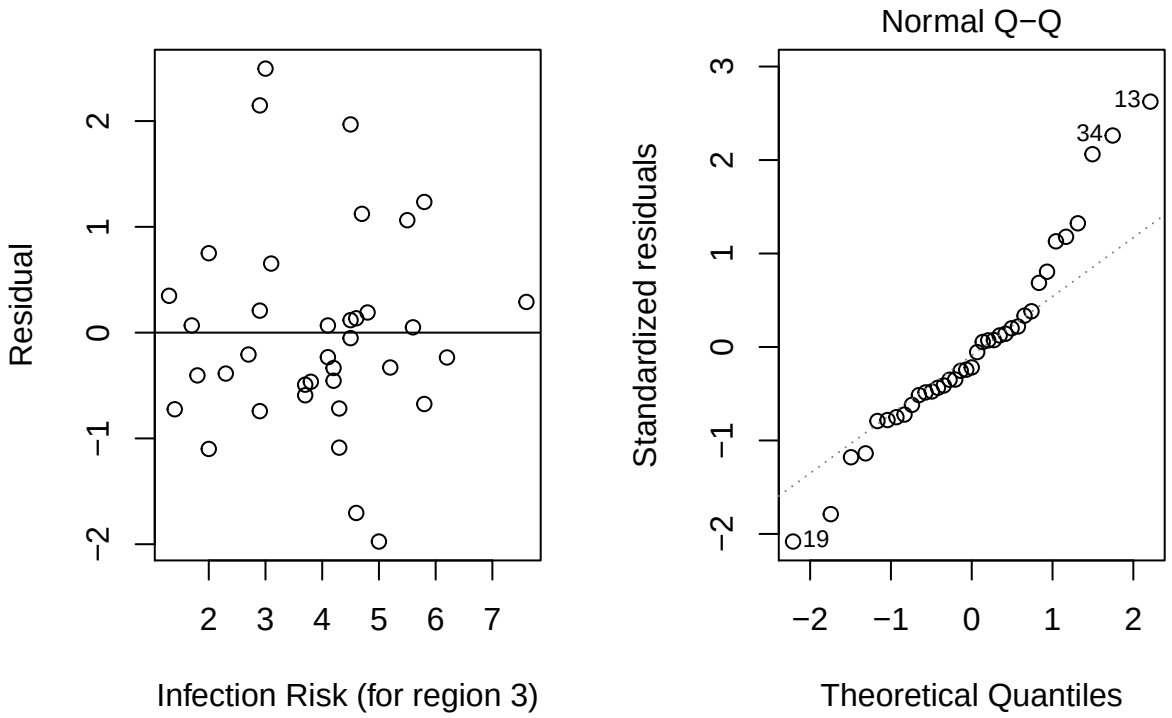
3.1.2 Region 2

```
par(mfrow = c(1,2))
plot(senic[senic$region==2,]$infec_risk, resid(models[[2]]),
     xlab ="Infection Risk (for region 2)",ylab = "Residual"); abline(h=0) # region 2
plot(models[[2]], which = 2)# normal probability plot
```



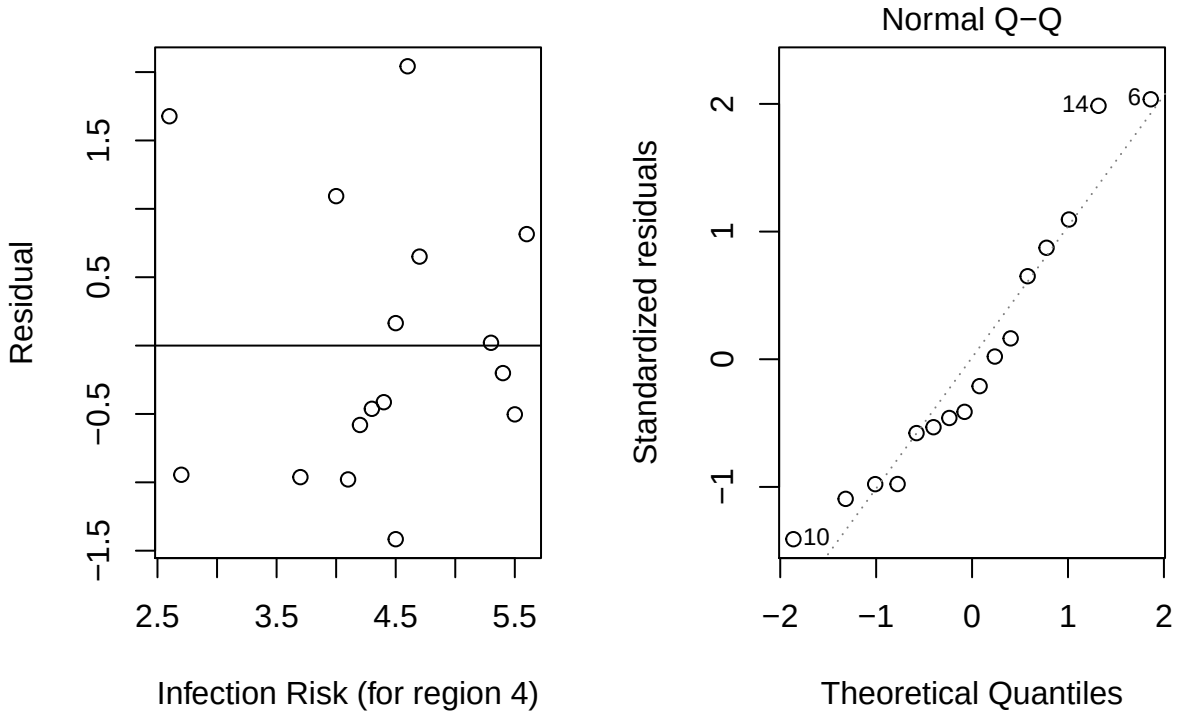
3.1.3 Region 3

```
par(mfrow = c(1,2))
plot(senic[senic$region==3,]$infec_risk, resid(models[[3]]),
     xlab ="Infection Risk (for region 3)",ylab = "Residual"); abline(h=0) # region 3
plot(models[[3]], which = 2)# normal probability plot
```



3.1.4 Region 4

```
par(mfrow = c(1,2))
plot(senic[senic$region==4,]$infec_risk, resid(models[[4]]),
     xlab ="Infection Risk (for region 4)",ylab = "Residual"); abline(h=0) # region 4
plot(models[[4]], which = 2)# normal probability plot
```



The residual versus predictor plots suggest that the error variance differ substantially across the four regions. There appear to be increasing error variance with levels of infection risk in region 1, while the other regions exhibit randomness.

Additionally, except region 2 there appear to be serious departure of error term distribution from normality.