Justin Yee
9220816444

STA108 A02: Applied Statistical Methods: Regression Analysis

Instructor: Amy Kim

I. Introduction

The goal is to develop the best multi-linear regression model to understand the correct relationship between Y and explanatory variables on the SENIC2 data. The primary objective is to use the SENIC2 data to explore the relationship between different variables and the length of stay at the hospital. The data involves the relationship between the length of stay in a hospital and ten different variables that were collected during the 1975-1976 period. To explore the data and their relationship, I'm using multi-linear regression to develop a statistically sound, interpretable model, that provides accurate predictions or insights.

II. Exploratory Analysis

Through these charts (SEE PLOTS & TABLES: *I - XIII*), we can see that there are a few outliers while each predictors seem to have a clear linear relationship with the response variable. We can also see that the response variable is approximately symmetric and normal with outliers in it. I will note these outliers' existence and investigate further to determine if they need to be removed in my next few steps.

III. Model Selection

Before beginning with my model selection procedures, I detected multicollinearity between each predictor variable to ensure that there aren't any predictors that correlate highly with each other that might end up being in my final model. Including irrelevant predictors in my final model could lead to increased difficulty of interpretation and boosted R^2. In this table (SEE PLOTS & TABLES: XIV, XI), we can see the VIF (Variance Inflation Factor) is extremely high with the predictors x5 (number of beds), x8 (average daily census), and x9 (number of nurses). I can see that x5 (number of beds) and x8 (average daily census) highly correlate with each other, so I investigate further about their relationship. I fit the full model without each predictor to find out which predictor contributes more to the model, that being x8. So I decided to remove x5 (number of beds) and x9 (number of nurses) from the full model to proceed with model selection. Since I want to build a model for correctness, I decided to use forward selection using the BIC criterion to find the best model. This involved starting with an empty model and adding one predictor at a time to the empty model to find the predictor that contributes the most to the model based on the BIC criterion. The ending best model was a combination of x1 (Age), x2 (Infection Risk), x7 (Region), and x8 (Average Daily Census). This model has an adjusted R^2 of 0.529 and a p-value of less than 0.0001 meaning this model is fit and explains the data well (SEE PLOTS & TABLES: XVI, XVII).

IV. Model Diagnostics

I noted there to be some outliers earlier, now I will determine if they will be removed or not based on high leverage points, high influential points, and outliers detection.

High Leverage Points:

Using the high leverage threshold of 2p/n, the following data are flagged as high leverage points (SEE PLOTS & TABLES: XVIII). Even though they are high-leverage points, I must ensure that these points are also high-influential points. I will note them being high-leverage points and continue to the next step.

High Influential Points:

Using Cook Distance and a threshold of 4/n-p-1, the following data are flagged as high influential points (SEE PLOTS & TABLES: XIX). Since they are high-influential points, it means that they influence parts of the regression analysis and we would want to remove that. Therefore, I removed these data points from the data treating them as outliers.

Outliers:

By standardizing the residuals, I am able to see and flag outliers from the data, the following data points are flagged as outliers based on the t-value (SEE PLOTS & TABLE: XX). These data points were already flagged as high-influential points, so they're already removed from the data. After cleaning the data, I refit the model to see how well the model is and it is much better than before (SEE PLOTS & TABLE: XXI, XXII).

Assumption: Normality

Using the Shapiro-Wilks Test for normality, I tested the residual of the model and received a p-value of 0.6818. Since the p-value is significantly high, I can conclude that the data is normally distributed. Therefore, the normality assumption is satisfied.

Assumption: Homoscedasticity

Using the Fligner-Kileen Test for homoscedasticity, I fitted the model and separated the data into two halves. The first group contains data less than the median and the second group contains data greater than the median. The Fligner-Kileen Test returned a p-value of 0.5271suggesting that the variance across both groups is approximately equal. Therefore, the homogeneity of variance assumption is satisfied.

Assumption: Independence and Linearity

Through these charts, we can see that the data is independent and approximately linear through the scatterplot of each predictor vs. response and the residual plot. (SEE PLOTS & TABLES: I-XIII). Therefore, the independent assumption and linearity assumption are satisfied.

V. Analysis and Interpretation

Null Hypothesis: Beta_i = 0, i = 1, 2, 7, 8. The predictor does not contribute significantly to the model

Alternative Hypothesis: Beta_i != 0, i = 1, 2, 7, 8. The predictor contributes significantly to the model

After doing ANOVA on the model, it returned p-values for each predictor. From the ANOVA table, we can see that we can reject the null hypothesis for each predictor under alpha level 0.05. (SEE PLOTS & TABLES: XXIII)

I also constructed an overall/family-wise/simultaneous confidence interval that captures the true change for each predictor. Under an alpha level of 0.05, the confidence interval for each predictor is as follows (SEE PLOTS & TABLES: XXIV).

I am 95% confident that the true change in the length of stay at a hospital when holding all variables constant lies between -0.628 and 6.432.

I am 95% confident that the true change in the length of stay at a hospital when age increased by one unit holding all other variables constant lies between 0.026 and 0.148.

I am 95% confident that the true change in the length of stay at a hospital when infection risk increases by one unit holding all other variables constant lies between 0.326 and 0.773.

I am 95% confident that the true change in the length of stay at a hospital when the region is NC holding all other variables constant lies between -1.297 and 0.059.

I am 95% confident that the true change in the length of stay at a hospital when the region is S holding all other variables constant lies between -1.674 and -0.336.

I am 95% confident that the true change in the length of stay at a hospital when the region is W holding all other variables constant lies between -3.222 and -1.473.

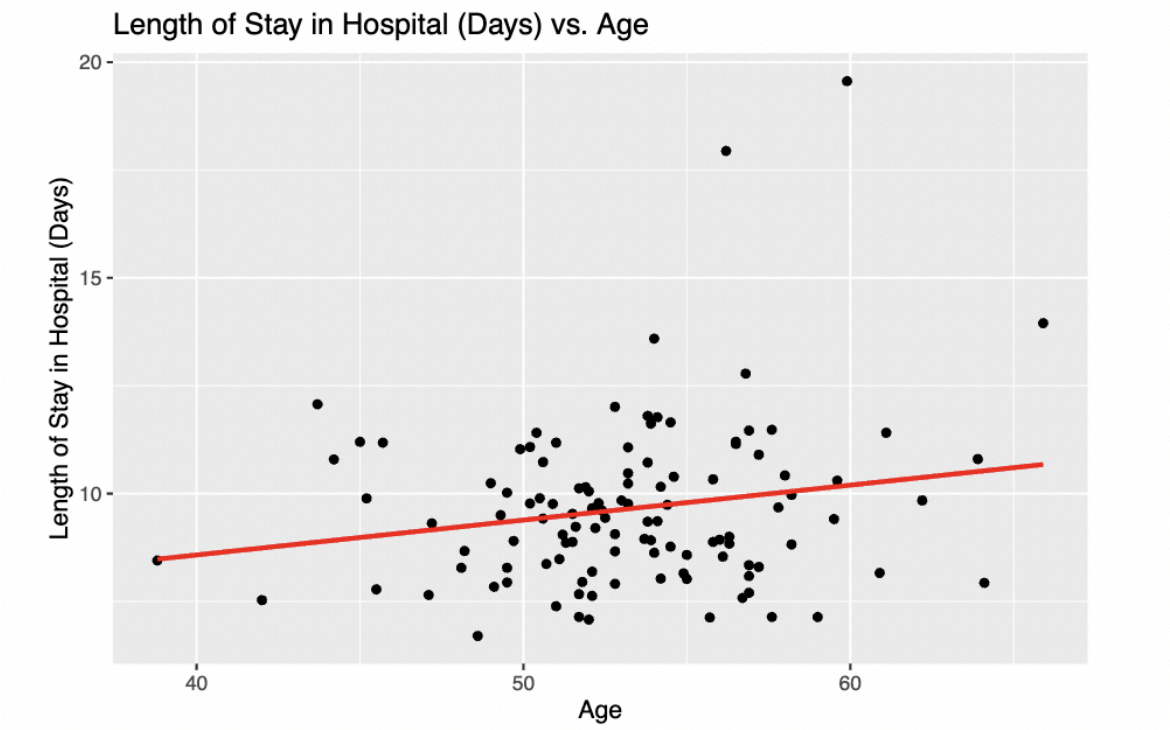
I am 95% confident that the true change in the length of stay at a hospital when the average daily census increases by one unit holding all other variables constant lies between 0.000031139 and 0.004007313.

VI. Conclusions

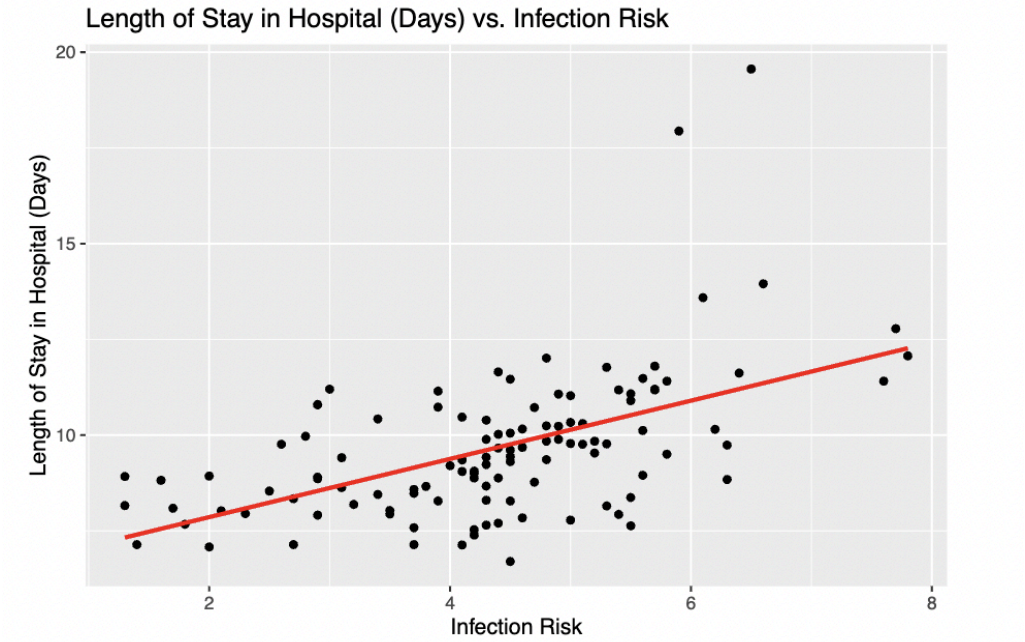
Based on the model selection, outliers removal, and model diagnostics, the valid multi-regression model provides accurate predictions or insights including four predictors; age, infection risk, region, and average daily census. This model satisfies all assumptions while each predictor contributes significantly to the model. However, one slight limitation of this model is that it doesn't consider the interaction effects between each predictor variable. One potential suggestion for this final model is to test individual interaction effects between each predictor. It is significant because they can reveal relationships in the data that would otherwise be hidden in models that only consider predictors independently.

Plots & Tables

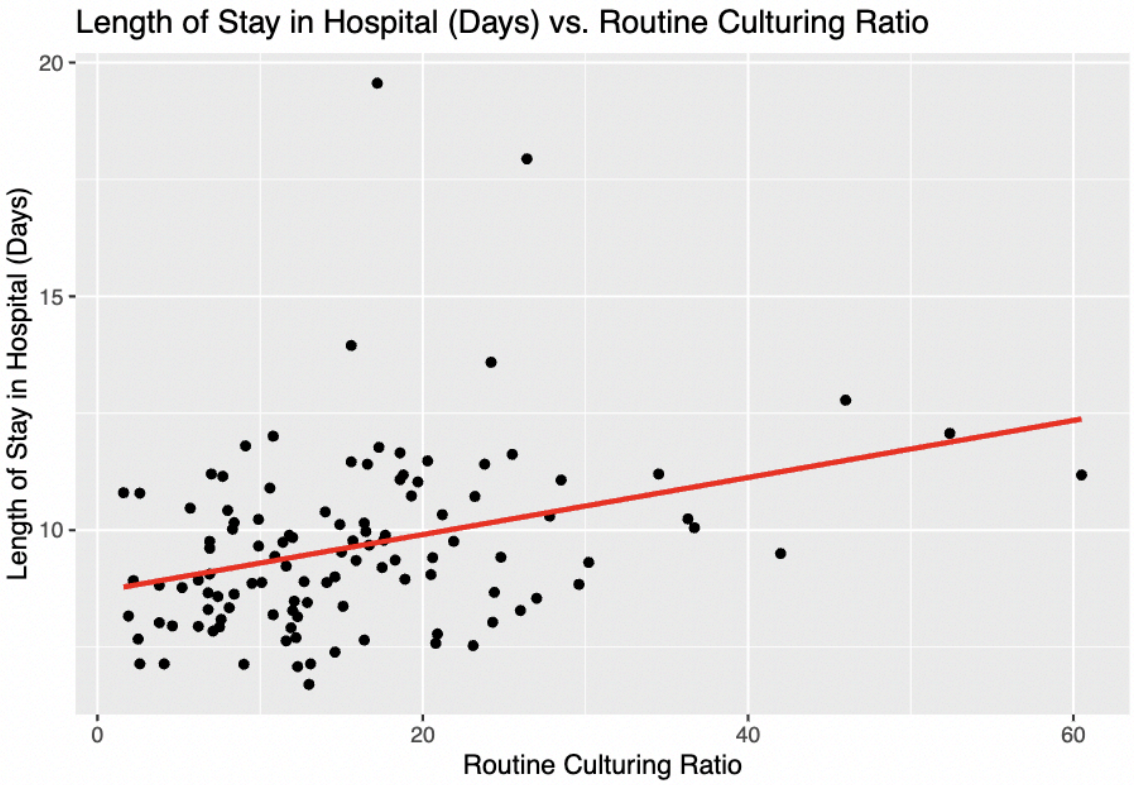
I.



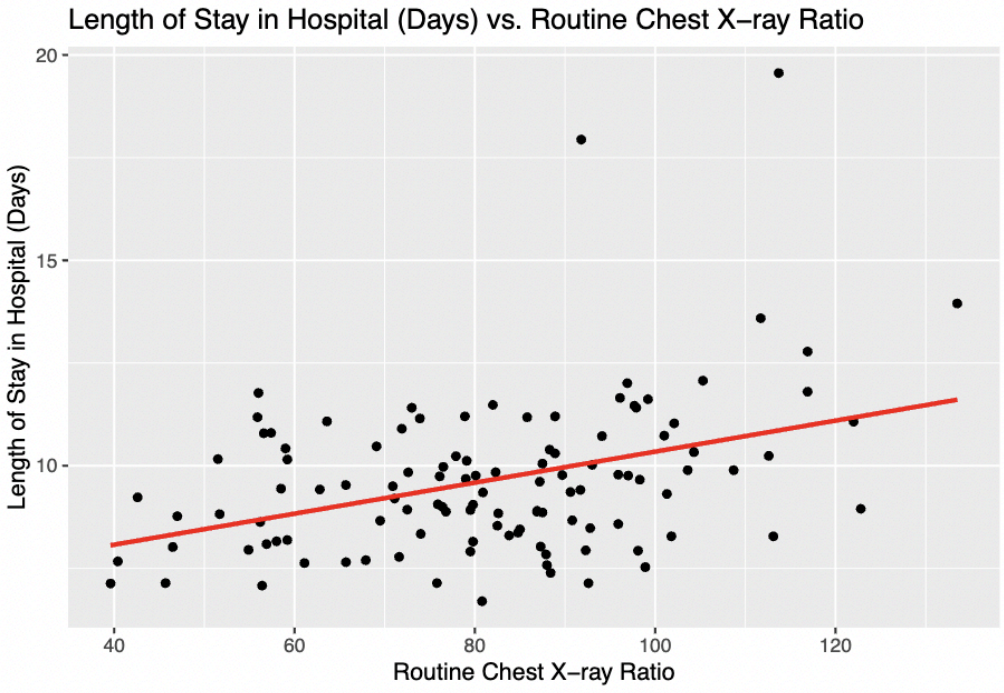
II.



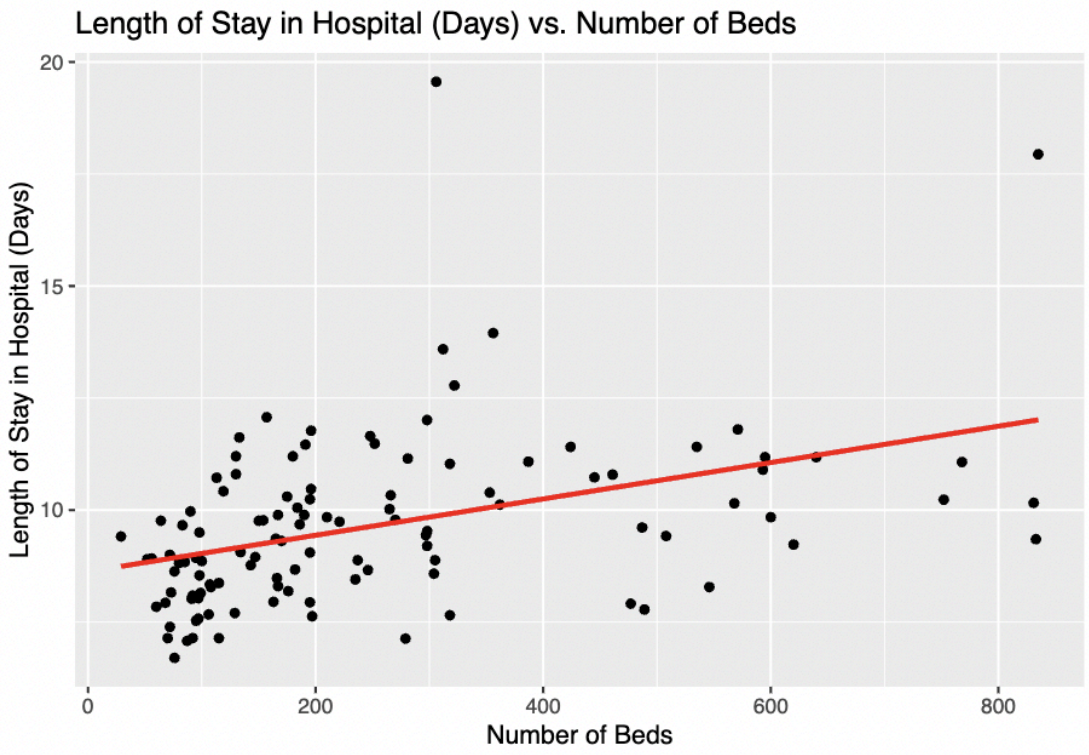
III.


Length of Stay in Hospital (Days) vs. Routine Culturing Ratio
Length of Stay in Hospital (Days)
20
15
10
0
20
40
60
Routine Culturing Ratio

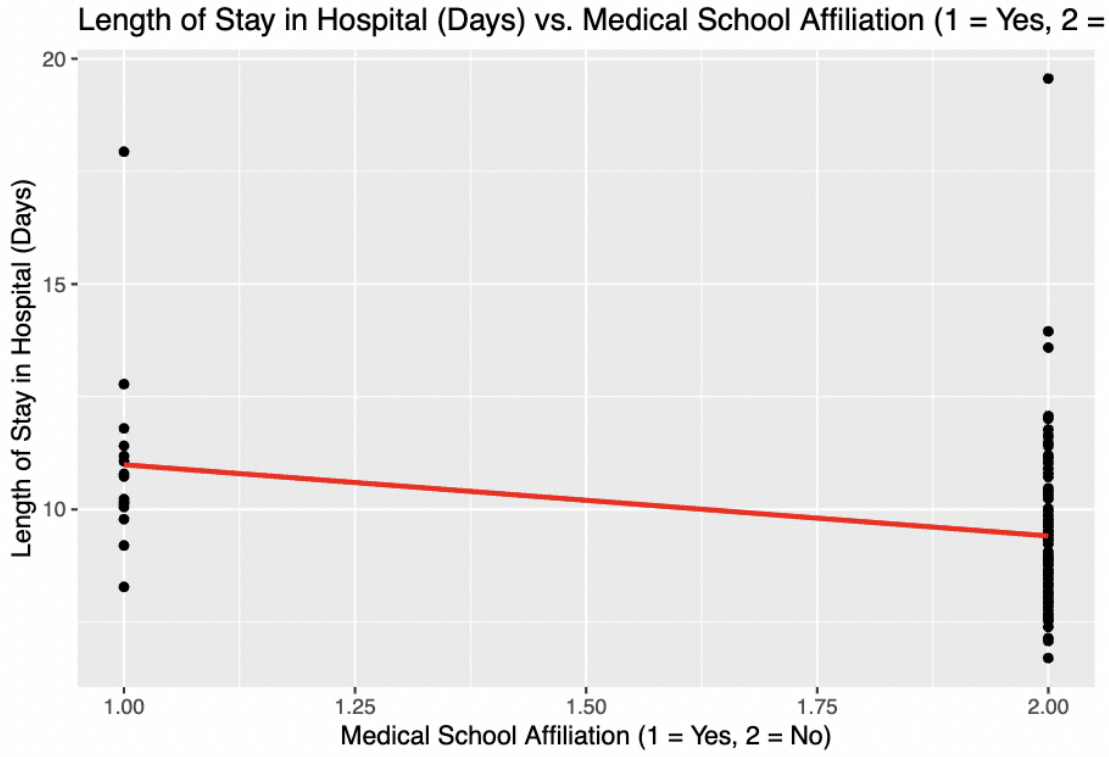
IV.


Length of Stay in Hospital (Days) vs. Routine Chest X−ray Ratio
Length of Stay in Hospital (Days)
20
15
10
40
60
80
100
120
Routine Chest X−ray Ratio

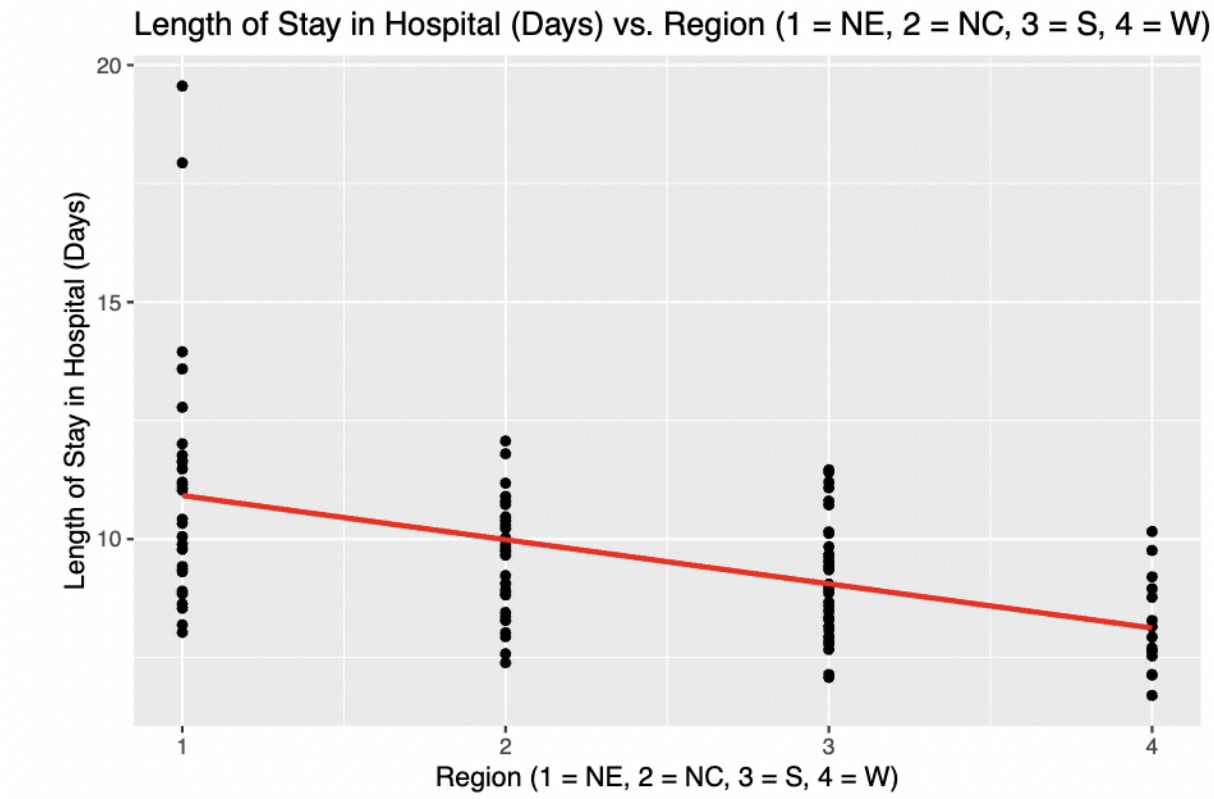
V.


Length of Stay in Hospital (Days) vs. Number of Beds
Length of Stay in Hospital (Days)
20
15
10
0
200
400
600
800
Number of Beds

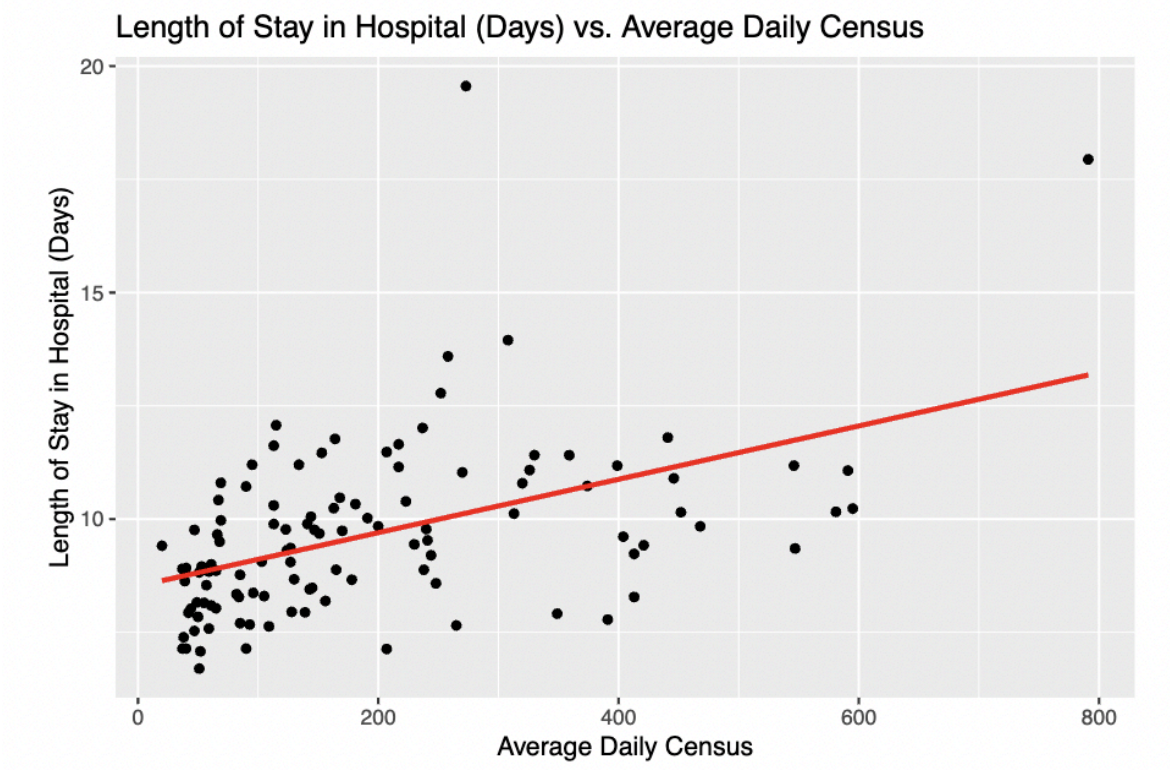
VI.


Length of Stay in Hospital (Days) vs. Medical School Affiliation (1 = Yes, 2 =
Length of Stay in Hospital (Days)
20
15
10
1.00
1.25
1.50
1.75
2.00
Medical School Affiliation (1 = Yes, 2 = No)

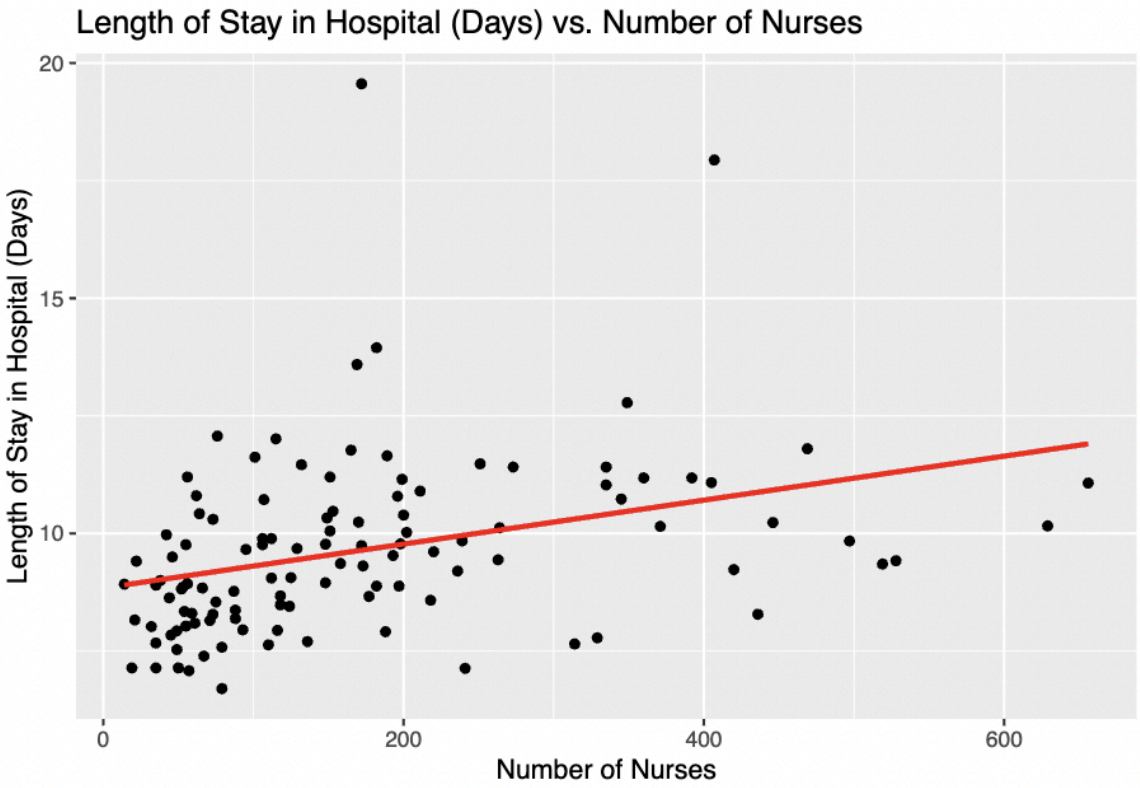
VII.


Length of Stay in Hospital (Days) vs. Region (1 = NE, 2 = NC, 3 = S, 4 = W)
Length of Stay in Hospital (Days)
20
15
10
1
2
3
4
Region (1 = NE, 2 = NC, 3 = S, 4 = W)

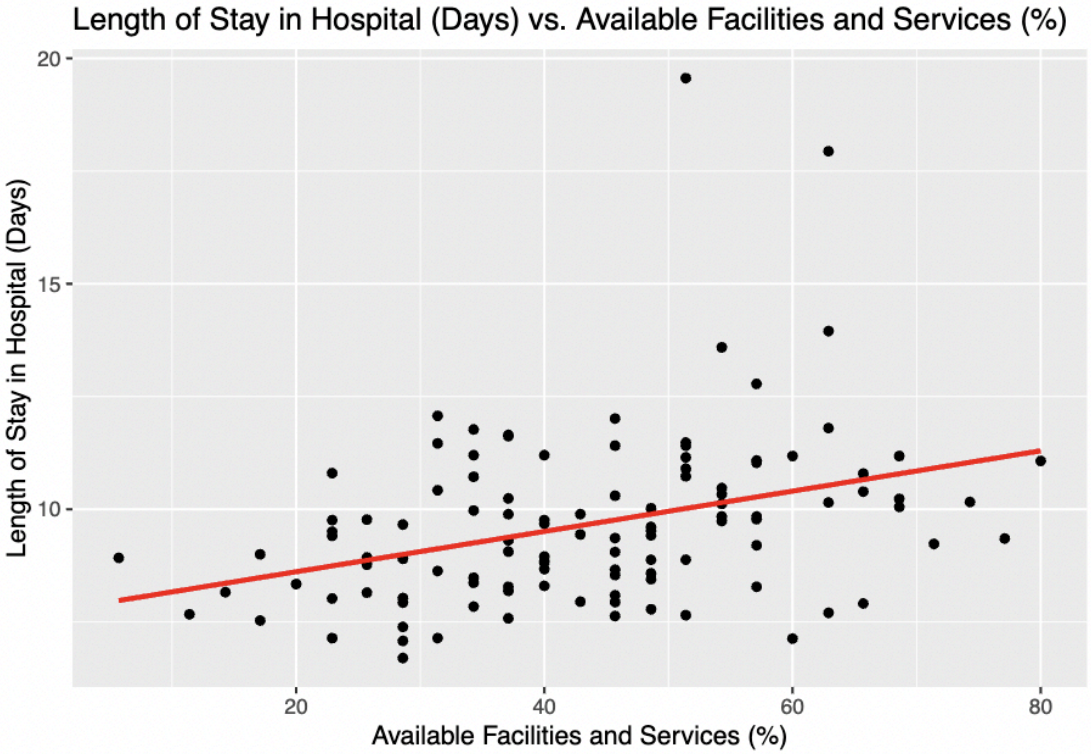
VIII.


Length of Stay in Hospital (Days) vs. Average Daily Census
Length of Stay in Hospital (Days)
20
15
10
0
200
400
600
800
Average Daily Census

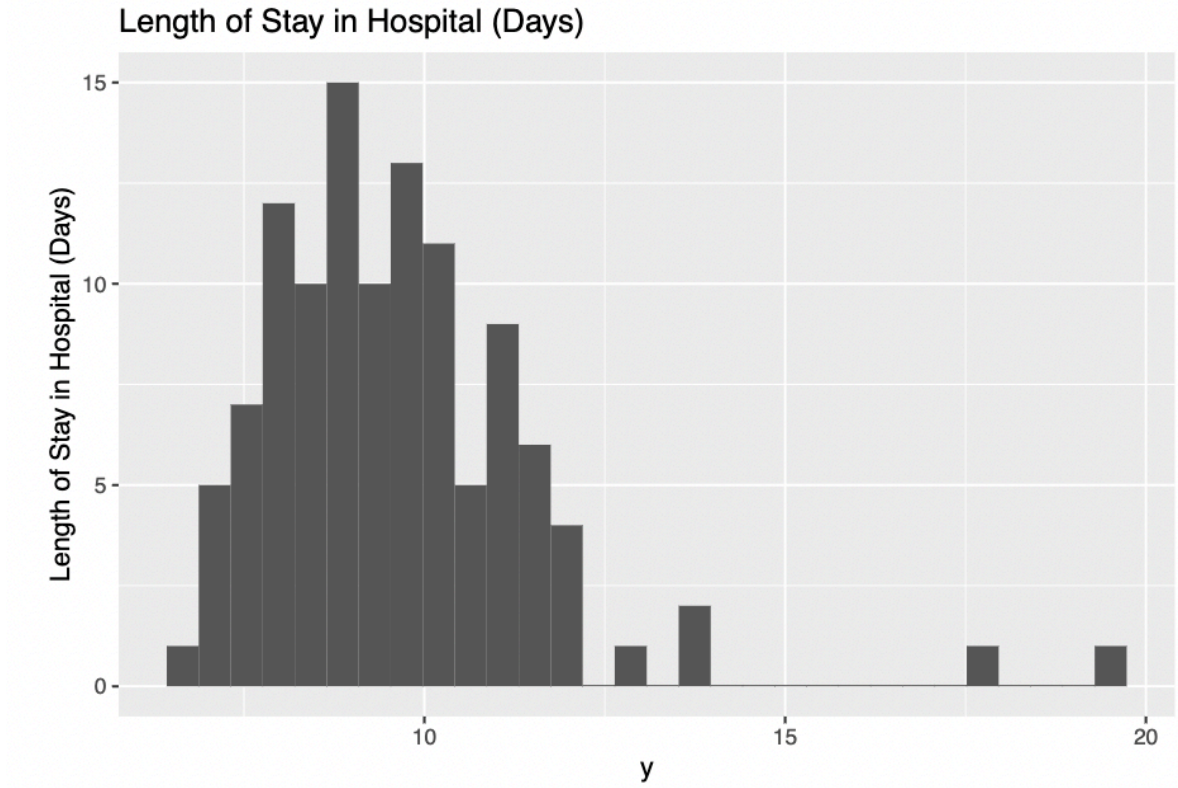
IX.


Length of Stay in Hospital (Days) vs. Number of Nurses
Length of Stay in Hospital (Days)
20
15
10
0
200
400
600
Number of Nurses

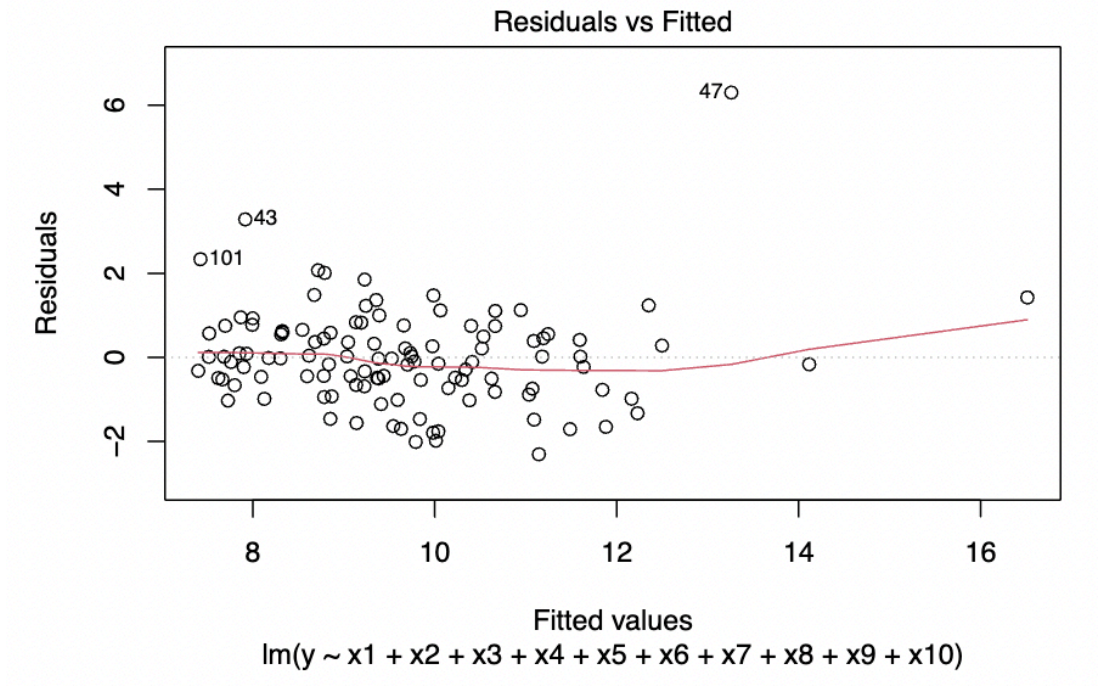
X.


Length of Stay in Hospital (Days) vs. Available Facilities and Services (%)
Length of Stay in Hospital (Days)
20
15
10
20
40
60
80
Available Facilities and Services (%)

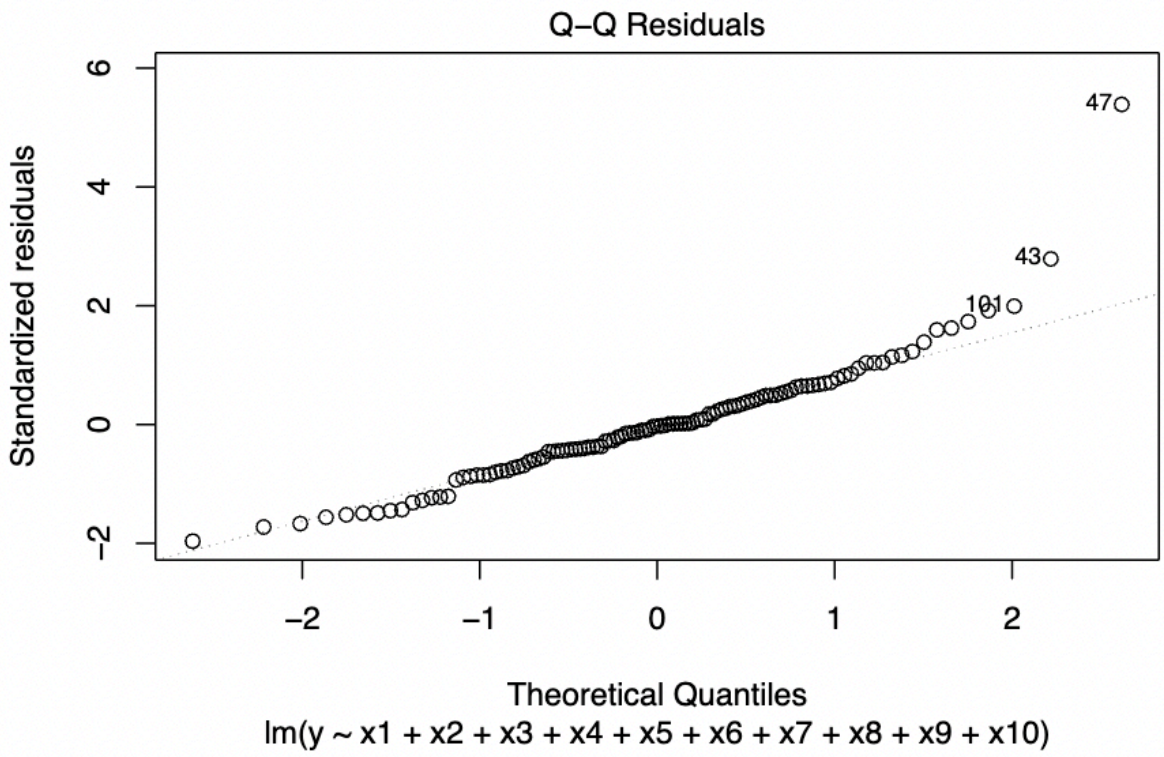
XI.


Length of Stay in Hospital (Days)
Length of Stay in Hospital (Days)
15
10
5
0
10
15
20
y

XII.


Residuals vs Fitted
Residuals
6
4
2
0
−2
47
43
101
8
10
12
14
16
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10)

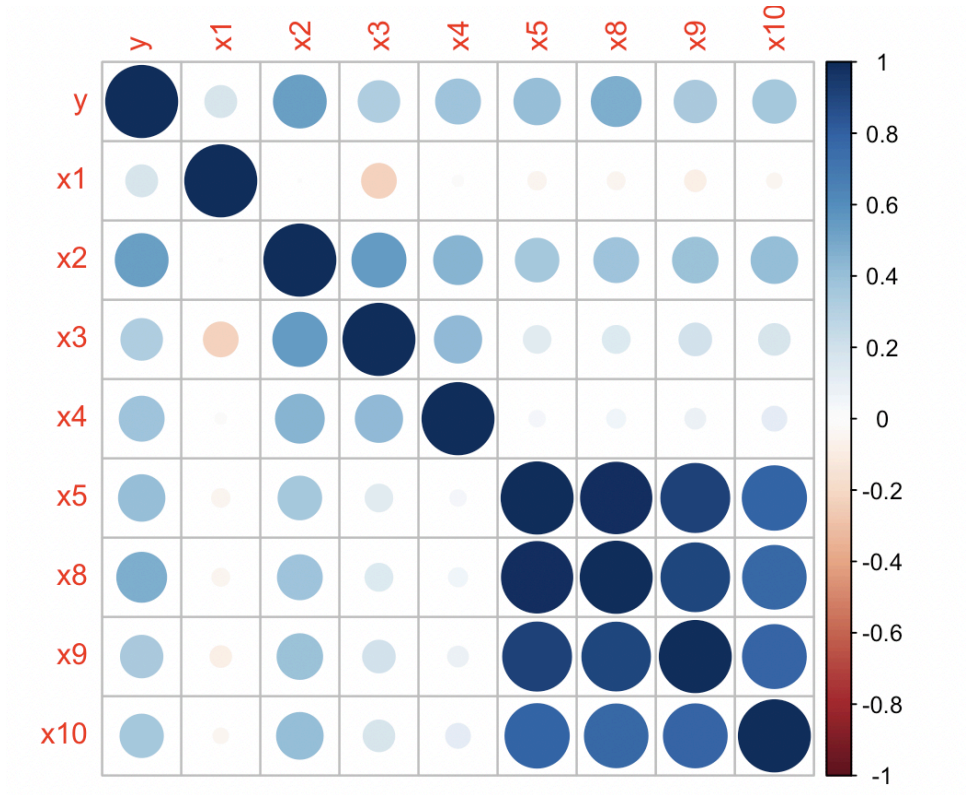
XIII.



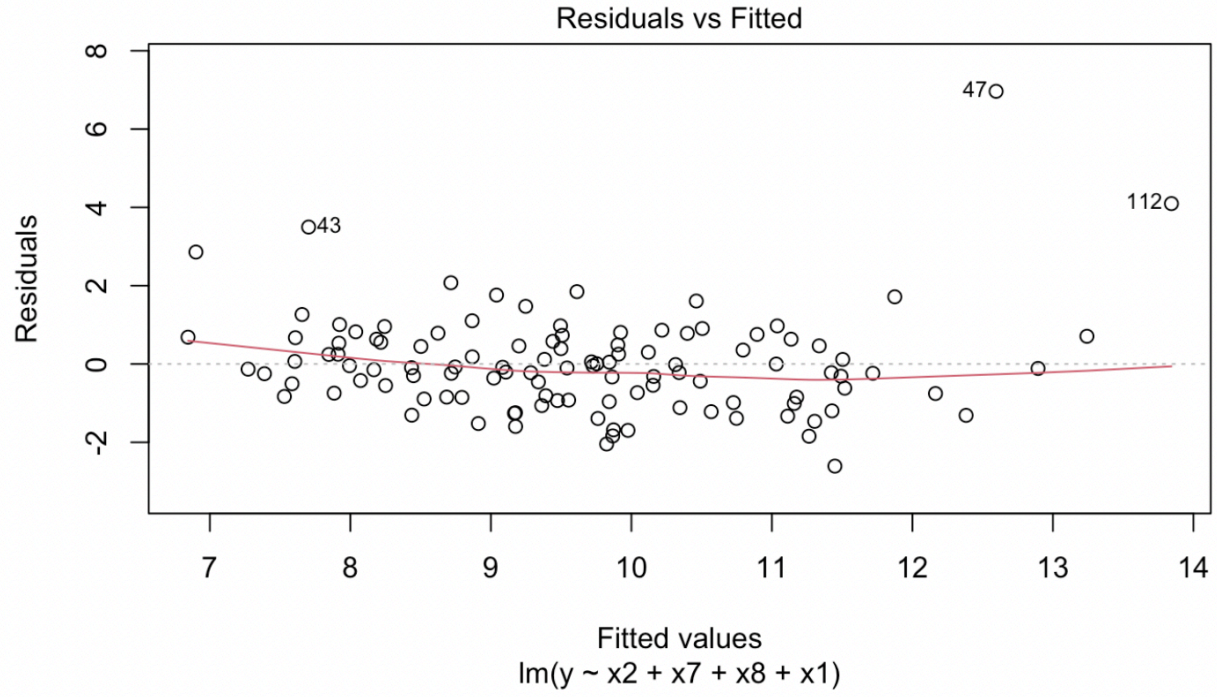
XIV.

	GVIF	Df	GVIF^(1/(2*Df))
x1	1.176172	1	1.084515
x2	2.154694	1	1.467888
x3	1.978520	1	1.406599
x4	1.416265	1	1.190070
x5	35.699204	1	5.974881
x6	1.855334	1	1.362107
x7	1.715222	3	1.094091
x8	34.211423	1	5.849053
x9	7.055523	1	2.656224
x10	3.241812	1	1.800503

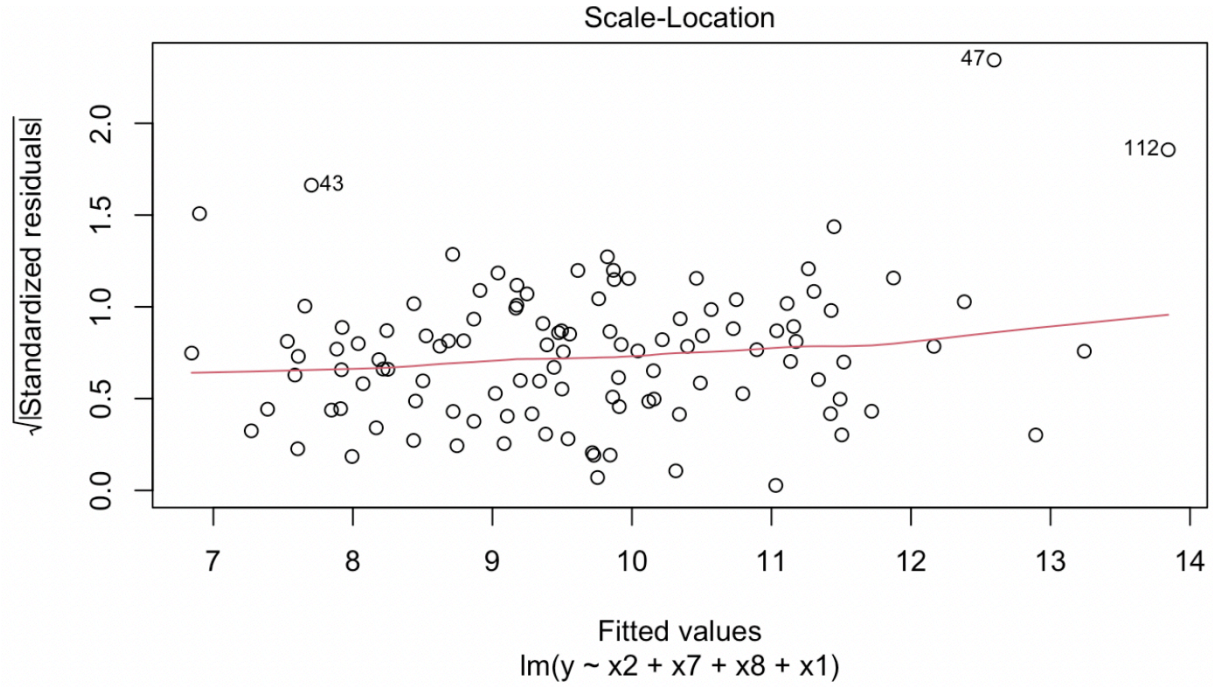
XV.


y
x1
x2
x3
x4
x5
x8
x9
x10
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

XVI.


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x2 + x7 + x8 + x1)
47
112
43
8
-2
0
2
4
6
8
7
8
9
10
11
12
13
14

XVII.



XVIII.

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
7.53	42	4.2	23.1	98.9	95	0	4	47	49	17.1
10.16	54.2	4.6	8.4	51.5	831	1	4	581	629	74.3
11.41	61.1	7.6	16.6	97.9	535	0	3	330	273	51.4
12.07	43.7	7.8	52.4	105.3	157	0	2	115	76	31.4
7.93	64.1	5.4	7.5	98.1	68	0	4	42	49	28.6
17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9

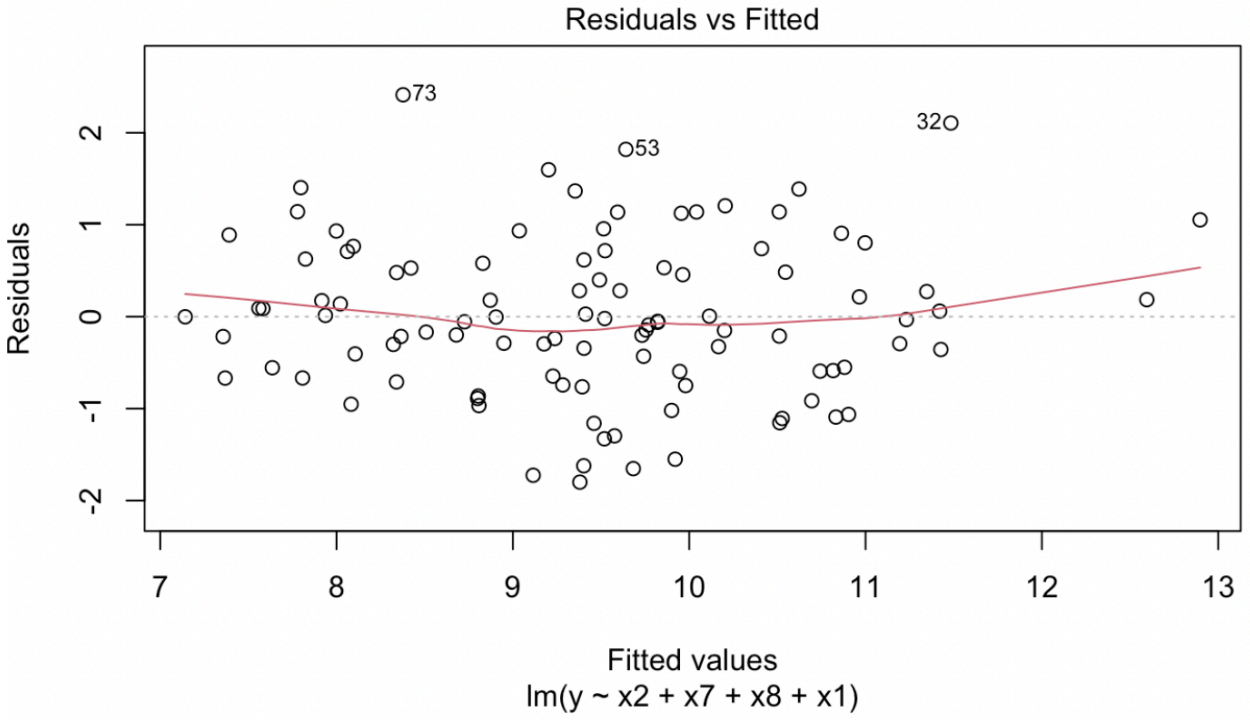
XIX.

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
8.84	56.3	6.3	29.6	82.6	85	0	1	59	66	40
11.2	45	3	7	78.9	130	0	3	95	56	34.3
19.56	59.9	6.5	17.2	113.7	306	0	1	273	172	51.4
12.07	43.7	7.8	52.4	105.3	157	0	2	115	76	31.4
9.76	53.2	2.6	6.9	80.1	64	0	4	47	55	22.9
17.94	56.2	5.9	26.4	91.8	835	1	1	791	407	62.9

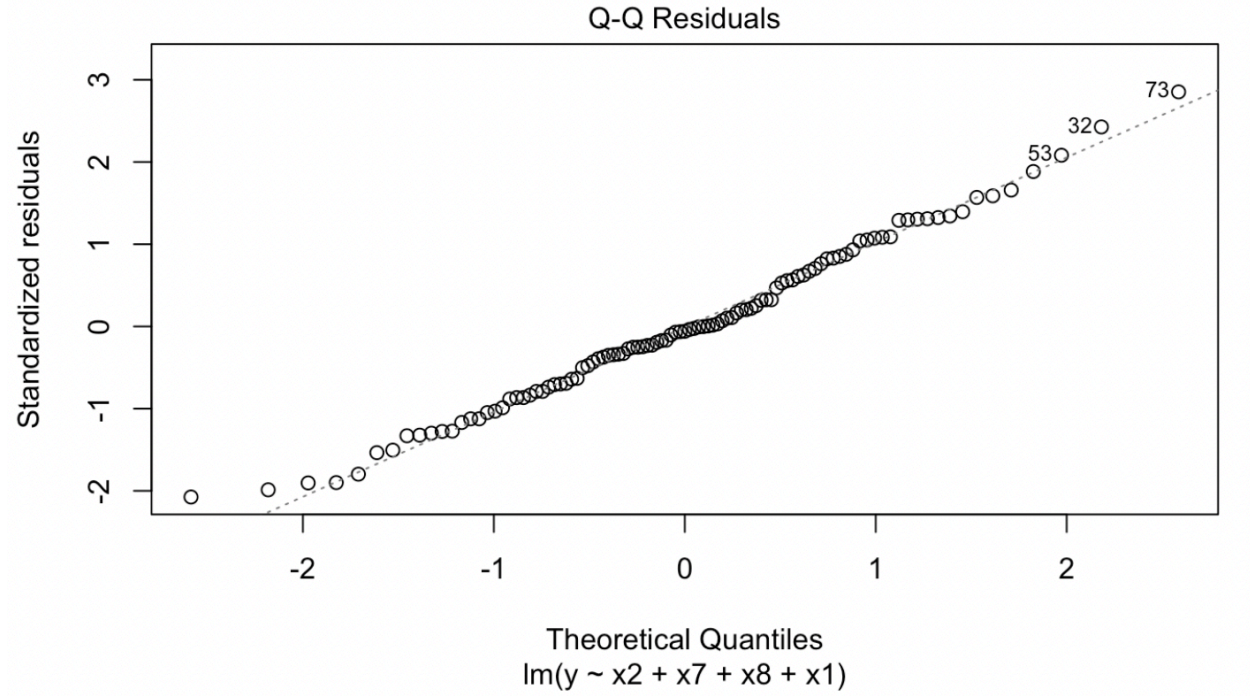
XX.

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
19.56	59.9	6.5	17.2	113.7	306	0	1	273	172	51.4

XXI.


Residuals vs Fitted
73
53
32
Residuals
2
1
0
-1
-2
7
8
9
10
11
12
13
Fitted values
lm(y ~ x2 + x7 + x8 + x1)

XXII.


Q-Q Residuals
73
32
53
Standardized residuals
3
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles
lm(y ~ x2 + x7 + x8 + x1)

XXIII.

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value              Pr(>F)
x2         1 72.697  72.697 91.3701 0.00000000000000132 ***
x7         3 52.989  17.663 22.1998 0.00000000005298711 ***
x8         1  4.336   4.336  5.4493           0.0216607 *
x1         1 12.221  12.221 15.3599           0.0001666 ***
Residuals 96 76.381   0.796
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

XXIV.

```
                  0.357 %     99.643 %
(Intercept) -0.627694025  6.431991874
x2           0.326131293  0.772836053
x72         -1.296832563  0.059060291
x73         -1.673988153 -0.335772103
x74         -3.222410709 -1.473052775
x8           0.000031139  0.004007313
x1           0.026036850  0.148367372
```

STA_Project 2 R Appendix

Justin Yee

2024-11-22

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:purrr':
##
##     some
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(EnvStats)
```

```
## Warning: package 'EnvStats' was built under R version 4.3.3
```

```
##
## Attaching package: 'EnvStats'
##
## The following object is masked from 'package:car':
##
##     qqPlot
##
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
library(onewaytests)
library(rcompanion)
```

```
## Warning: package 'rcompanion' was built under R version 4.3.3
```

```
## Registered S3 method overwritten by 'DescTools':
##   method        from
##   print.palette wesanderson
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
data = read_csv("/Users/justin/Documents/RStudio/SENIC2.csv")
```

```
## Rows: 113 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (11): y, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
options(scipen = 999)
```

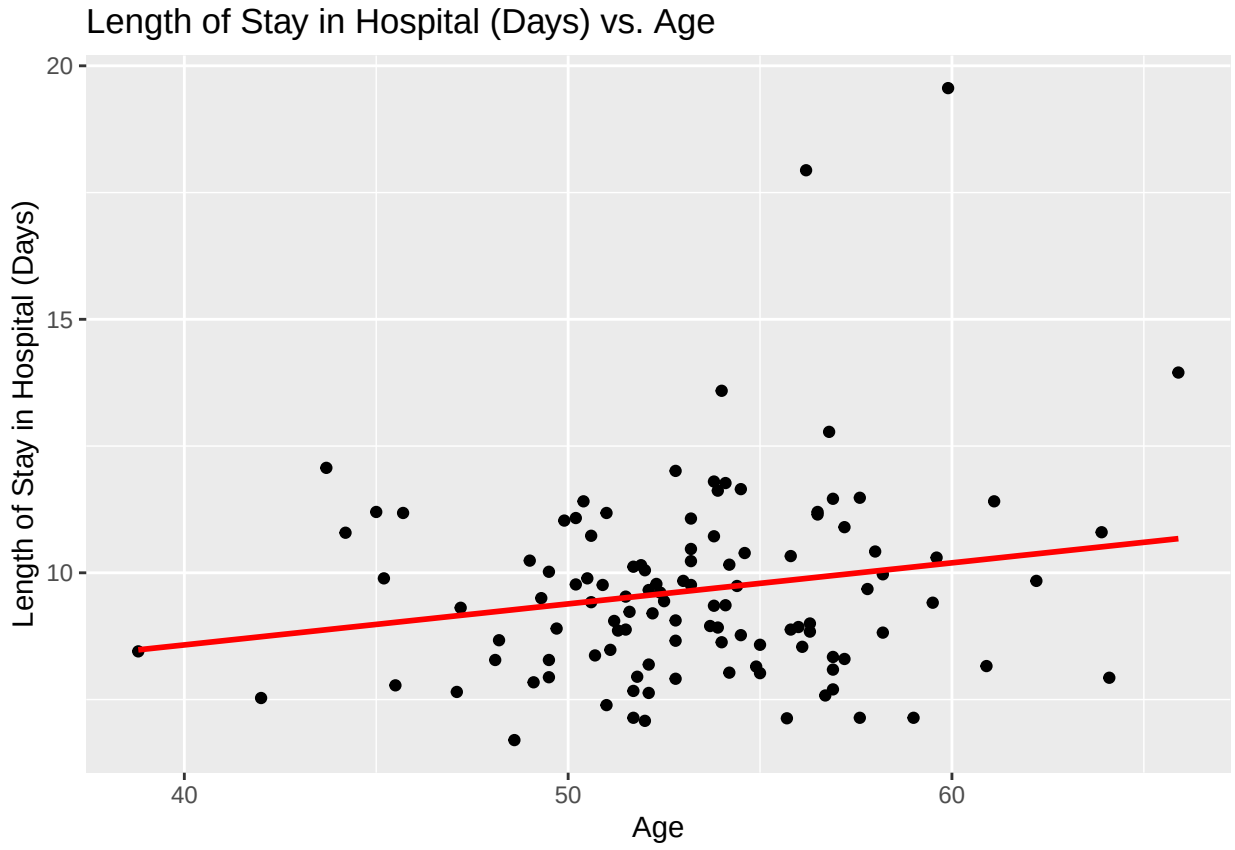
```
summary(data)
```

```
##        y               x1              x2              x3
##  Min.   : 6.700   Min.   :38.80   Min.   :1.300   Min.   : 1.60
##  1st Qu.: 8.340   1st Qu.:50.90   1st Qu.:3.700   1st Qu.: 8.40
##  Median : 9.420   Median :53.20   Median :4.400   Median :14.10
##  Mean   : 9.648   Mean   :53.23   Mean   :4.355   Mean   :15.79
##  3rd Qu.:10.470   3rd Qu.:56.20   3rd Qu.:5.200   3rd Qu.:20.30
##  Max.   :19.560   Max.   :65.90   Max.   :7.800   Max.   :60.50
```

```
##        x4               x5              x6            x7
##  Min.   : 39.60   Min.   : 29.0   Min.   :1.00   Min.   :1.000
##  1st Qu.: 69.50   1st Qu.:106.0   1st Qu.:2.00   1st Qu.:2.000
##  Median : 82.30   Median :186.0   Median :2.00   Median :2.000
##  Mean   : 81.63   Mean   :252.2   Mean   :1.85   Mean   :2.363
##  3rd Qu.: 94.10   3rd Qu.:312.0   3rd Qu.:2.00   3rd Qu.:3.000
##  Max.   :133.50   Max.   :835.0   Max.   :2.00   Max.   :4.000
##        x8              x9             x10
##  Min.   : 20.0   Min.   : 14.0   Min.   : 5.70
##  1st Qu.: 68.0   1st Qu.: 66.0   1st Qu.:31.40
##  Median :143.0   Median :132.0   Median :42.90
##  Mean   :191.4   Mean   :173.2   Mean   :43.16
##  3rd Qu.:252.0   3rd Qu.:218.0   3rd Qu.:54.30
##  Max.   :791.0   Max.   :656.0   Max.   :80.00
```

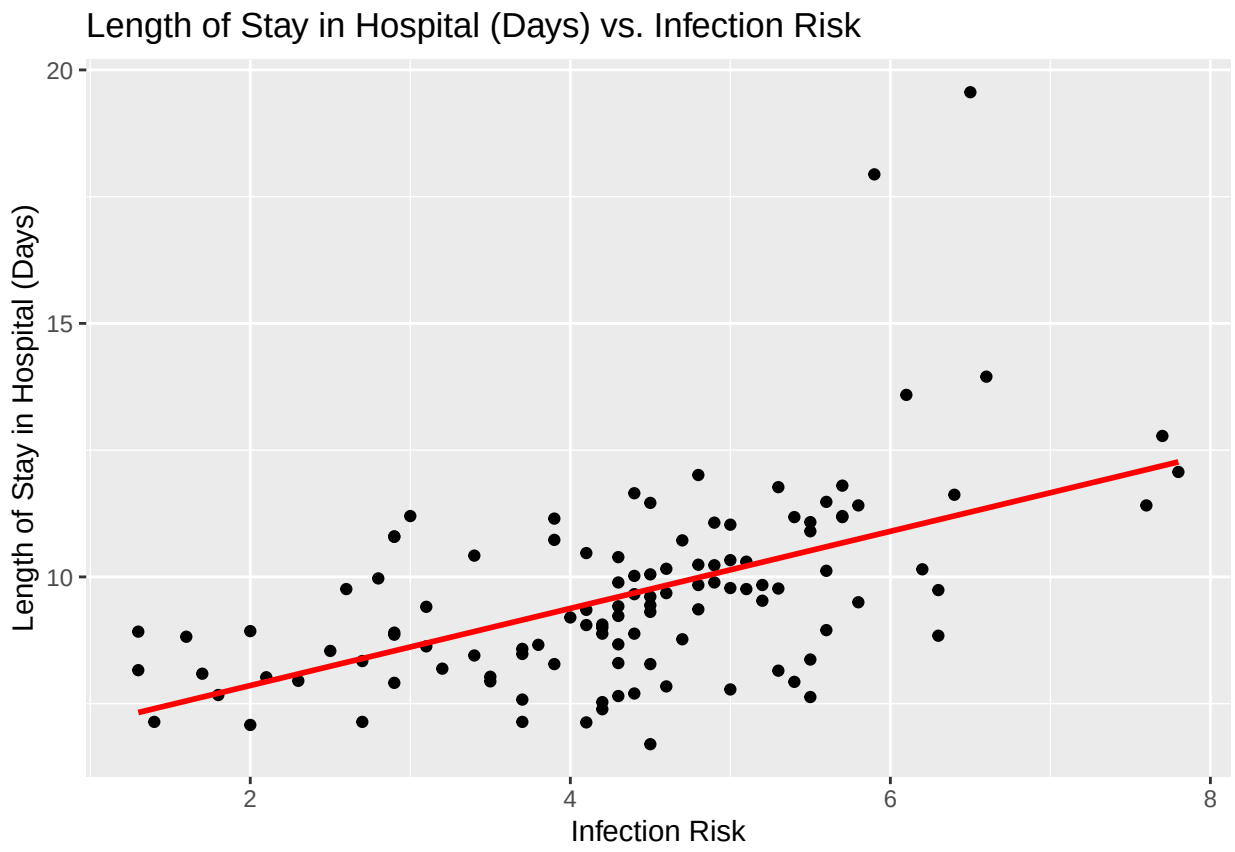
```
ggplot(data, aes(x = x1, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Age", x = "Age",
       y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



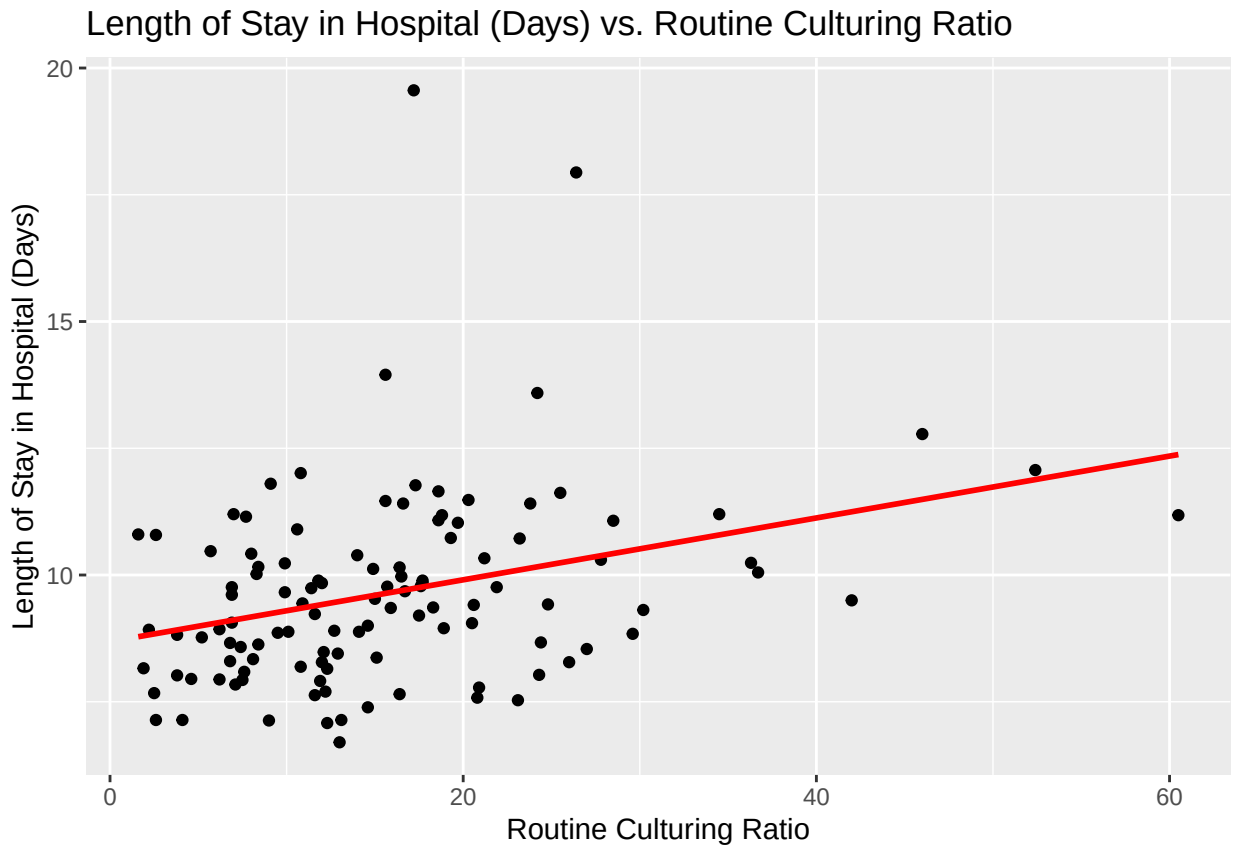
```
ggplot(data, aes(x = x2, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Infection Risk", x = "Infection Risk",
       y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



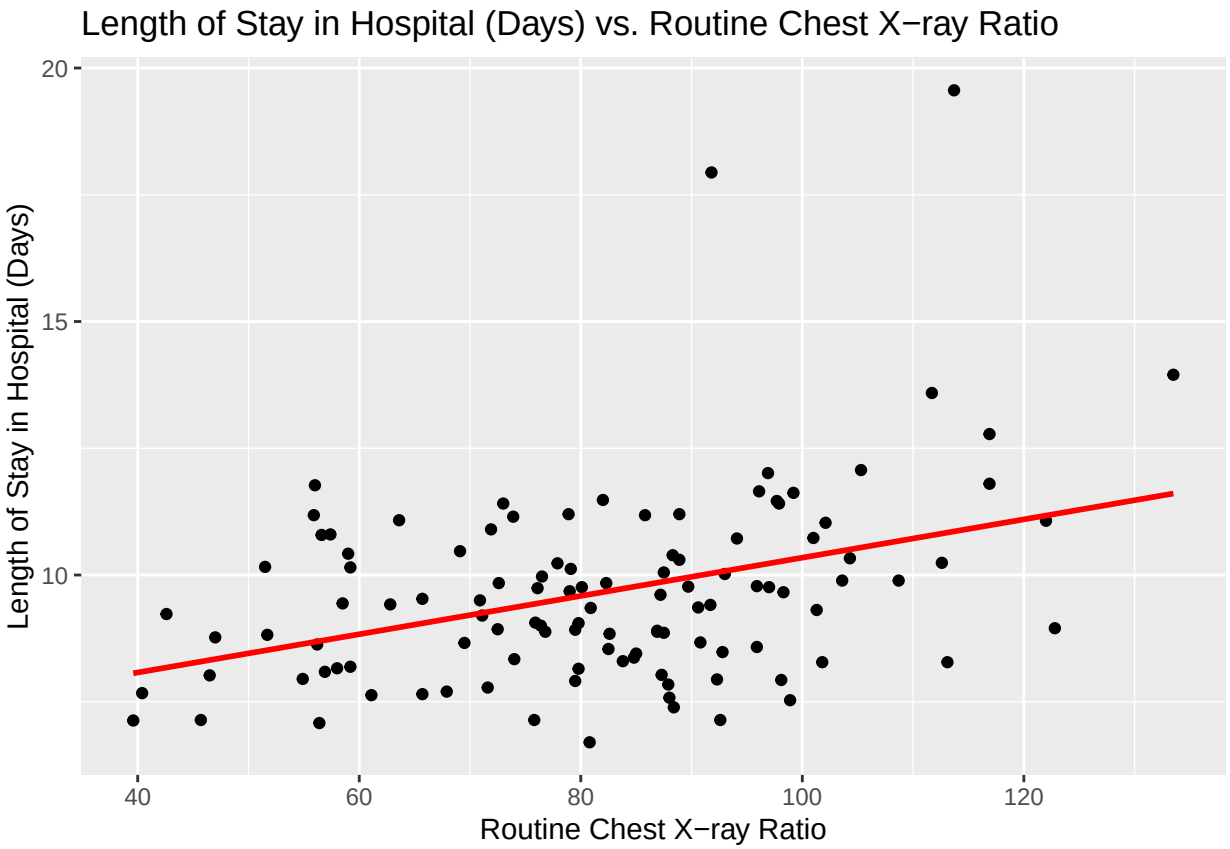

```
ggplot(data, aes(x = x3, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Routine Culturing Ratio",
       x = "Routine Culturing Ratio", y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



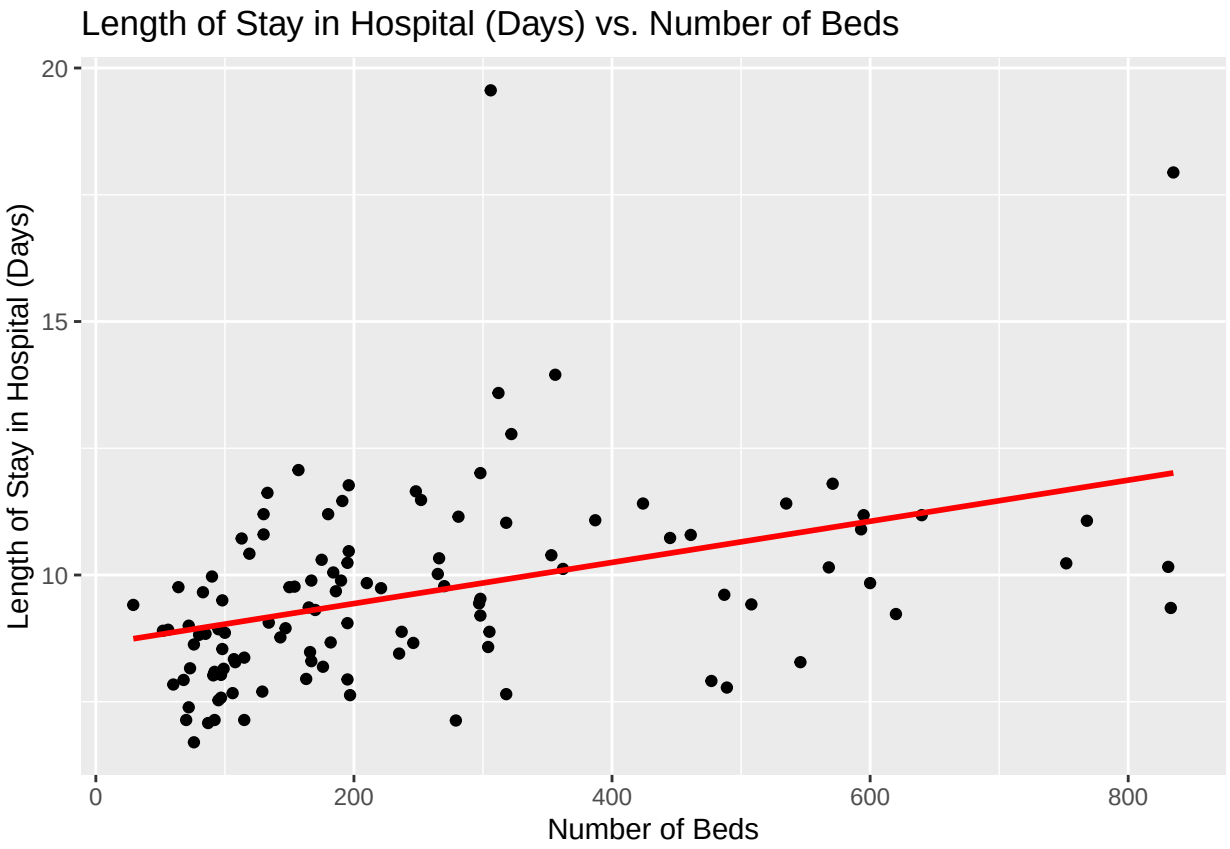

```
ggplot(data, aes(x = x4, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Routine Chest X-ray Ratio",
       x = "Routine Chest X-ray Ratio", y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



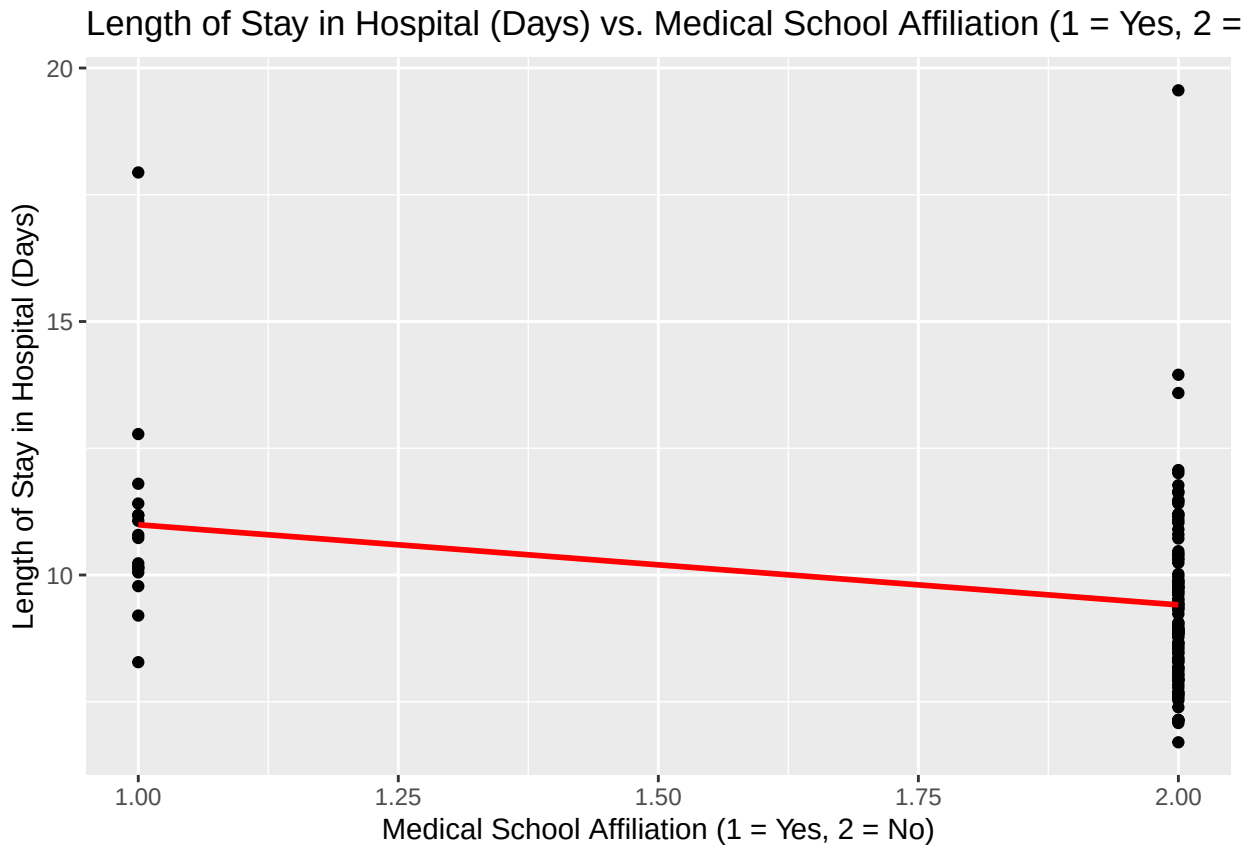

```
ggplot(data, aes(x = x5, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Number of Beds", x = "Number of Beds",
       y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



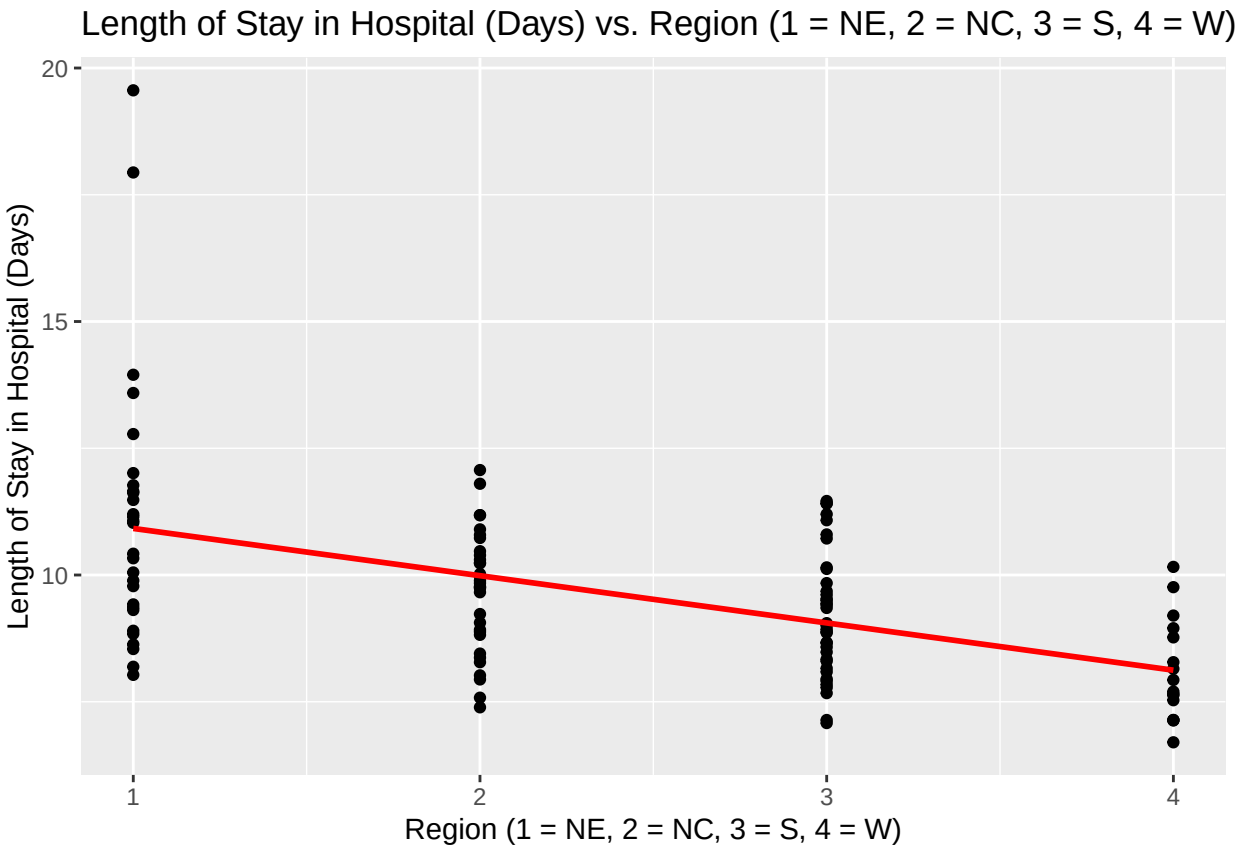

```
ggplot(data, aes(x = x6, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Medical School Affiliation (1 = Yes, 2 = No)",
       x = "Medical School Affiliation (1 = Yes, 2 = No)", y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



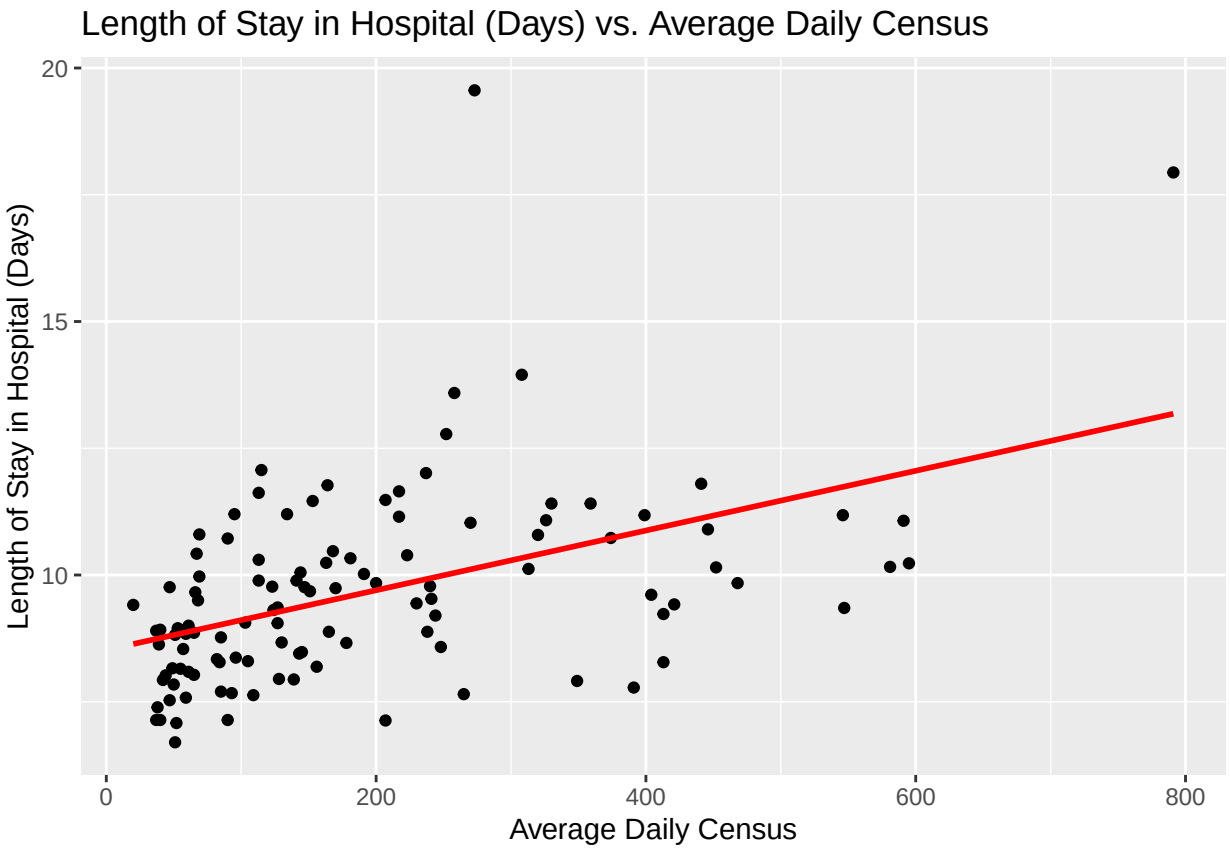

```
ggplot(data, aes(x = x7, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Region (1 = NE, 2 = NC, 3 = S, 4 = W)",
       x = "Region (1 = NE, 2 = NC, 3 = S, 4 = W)", y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



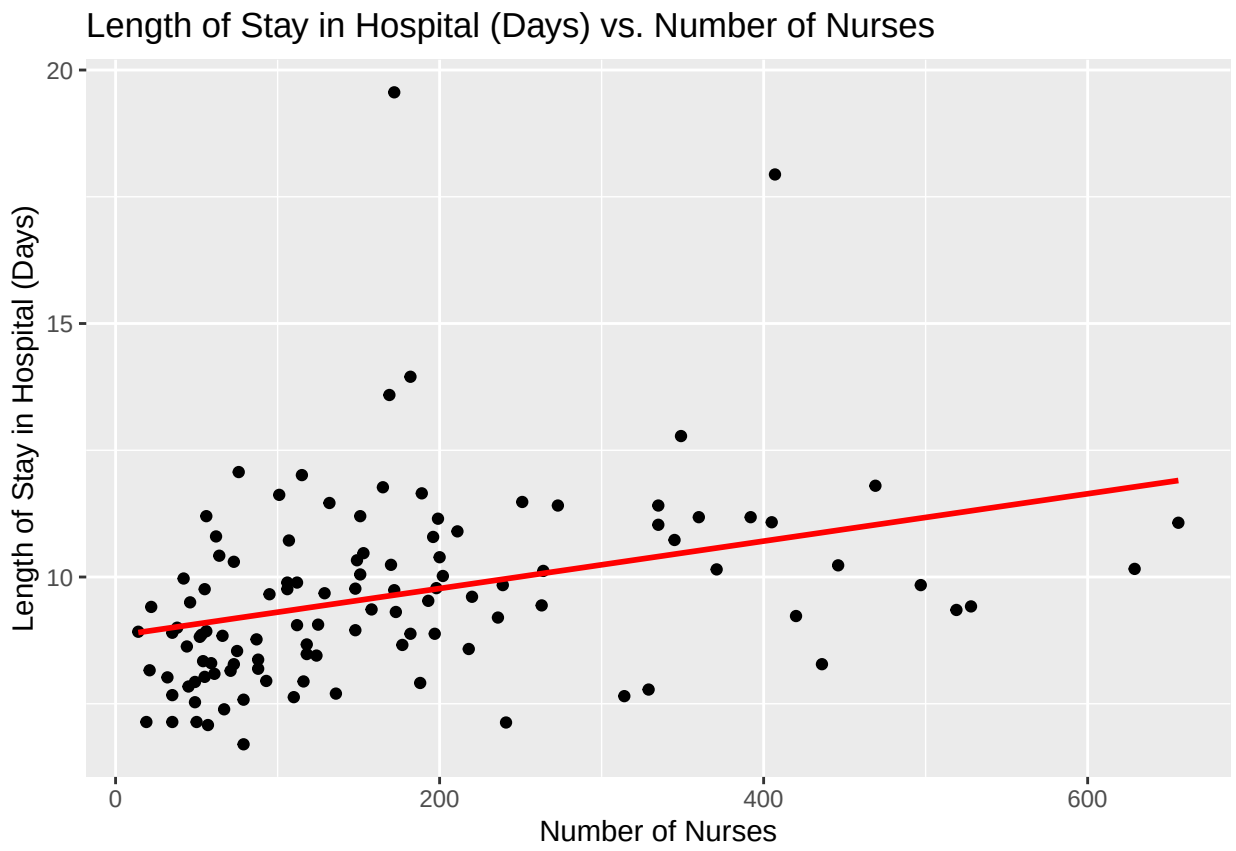

```
ggplot(data, aes(x = x8, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Average Daily Census",
       x = "Average Daily Census", y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



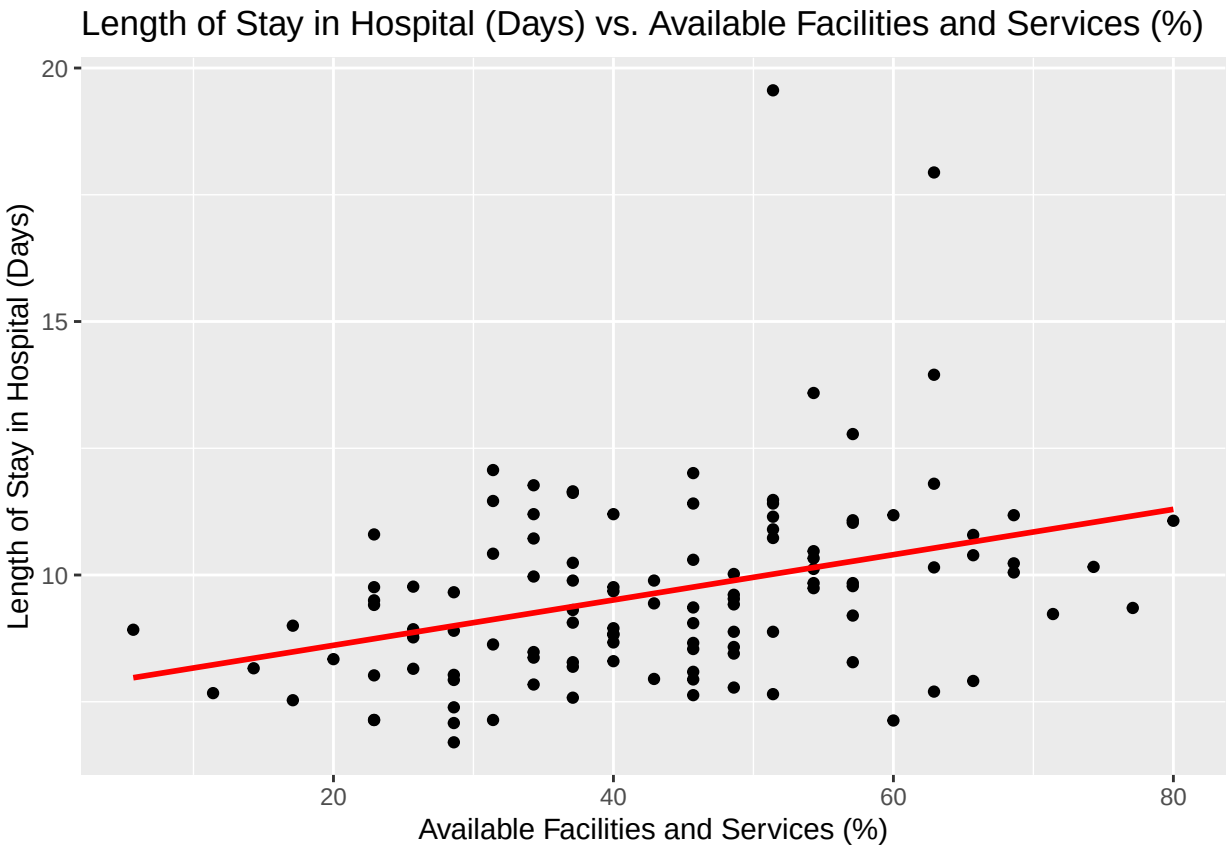

```
ggplot(data, aes(x = x9, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Number of Nurses", x = "Number of Nurses",
       y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



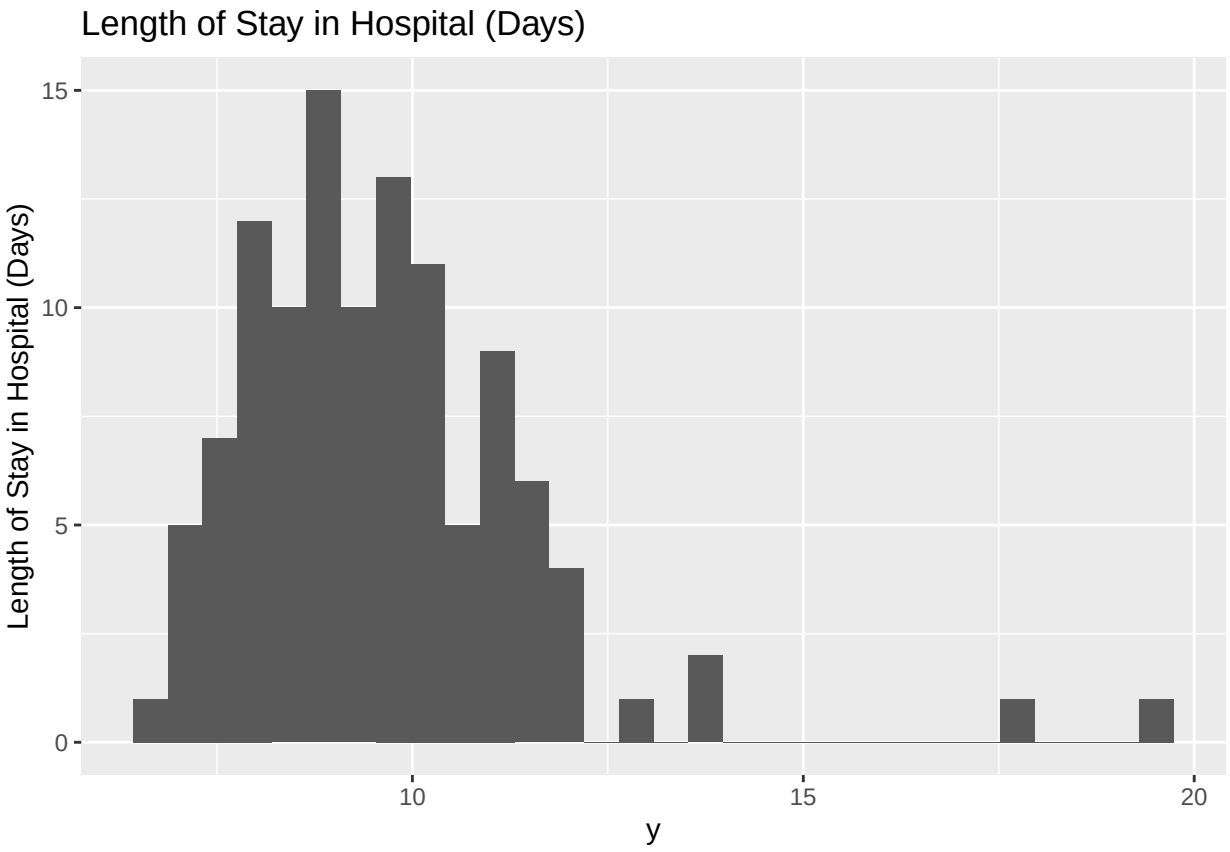

```
ggplot(data, aes(x = x10, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Length of Stay in Hospital (Days) vs. Available Facilities and Services (%)",
       x = "Available Facilities and Services (%)", y = "Length of Stay in Hospital (Days)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

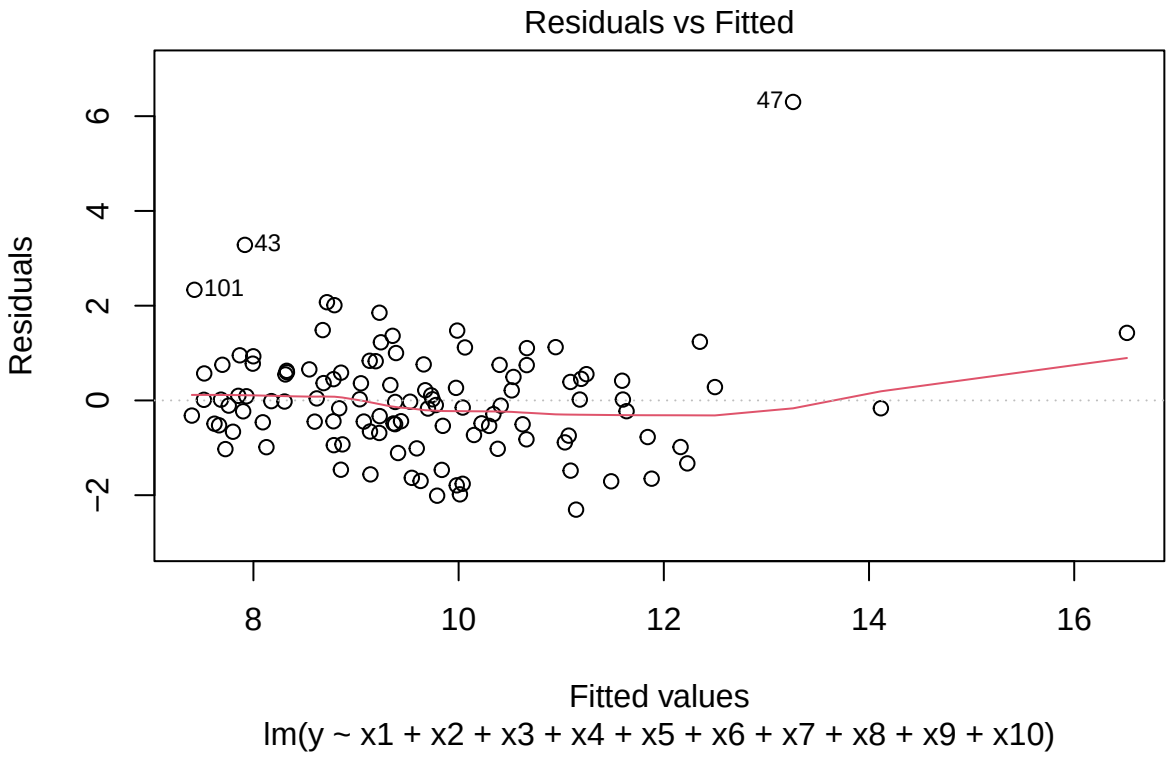


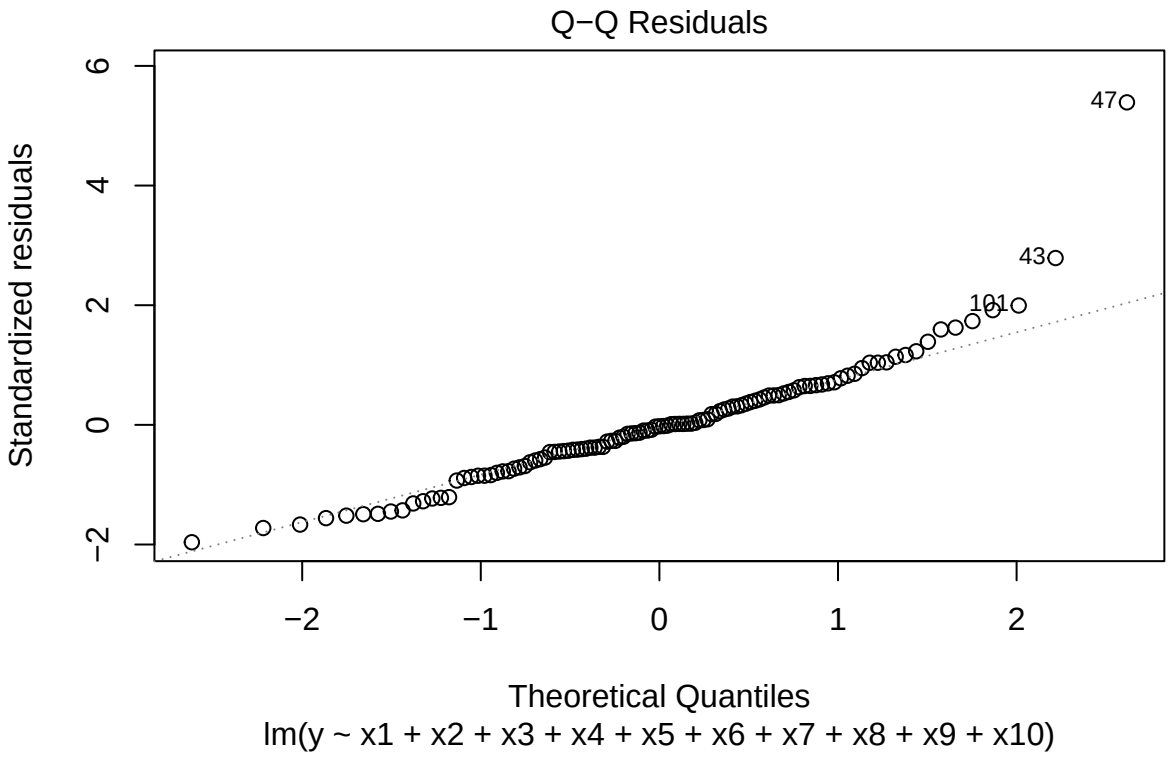

```
ggplot(data, aes(x = y)) +
  geom_histogram() +
  labs(title = "Length of Stay in Hospital (Days)", y = "Length of Stay in Hospital (Days)")
```

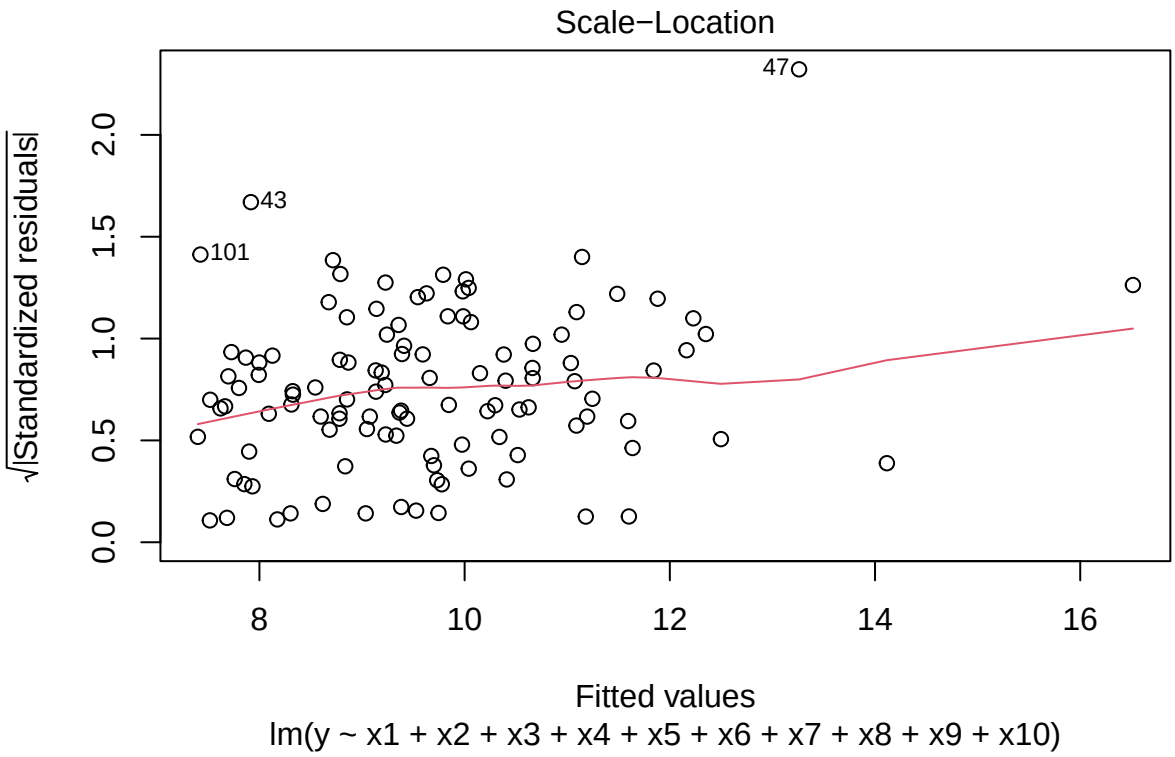
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

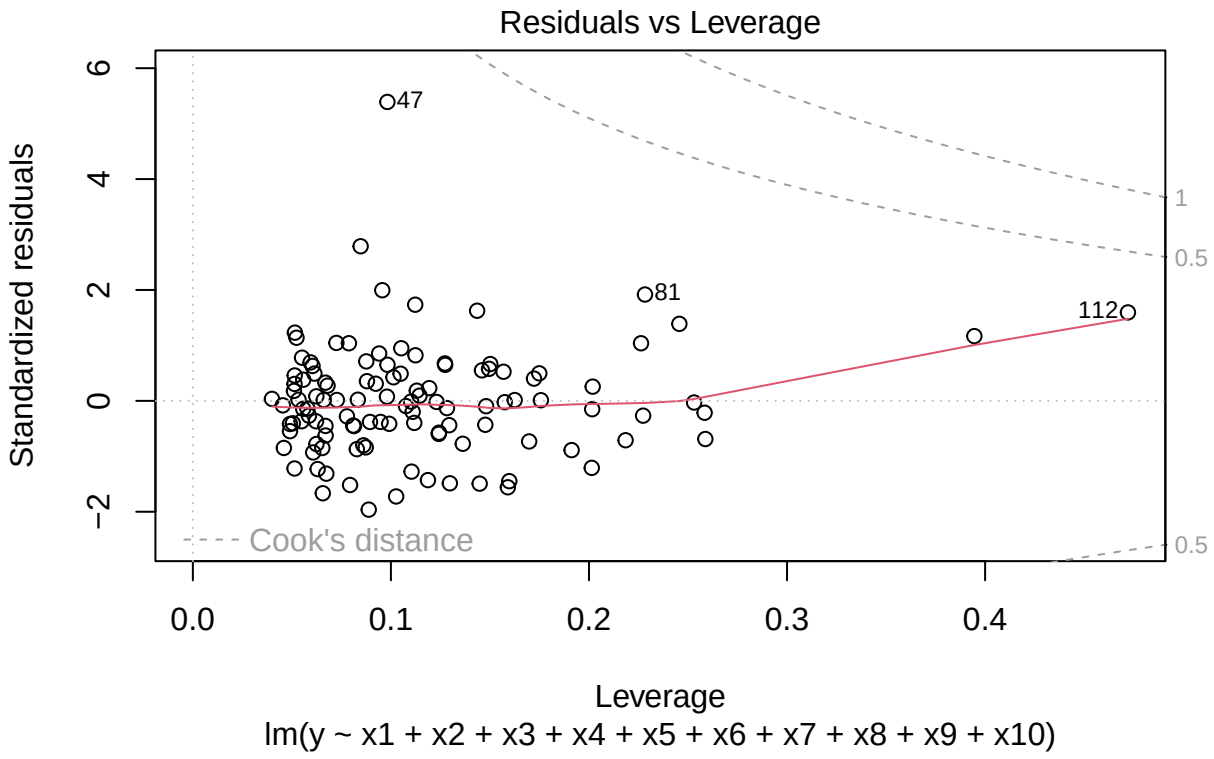



```
#There are a few outliers while each predictors seem to have a linear relationship
#with the response variable. Though these charts, we can see that there is a clear
#linear relationship. I want to find the model that best represents correctness.

data$x6 <- factor(data$x6, levels = c(1, 2), labels = c(1, 0))
data$x7 = factor(data$x7, levels = c(1, 2, 3, 4), labels = c(1, 2, 3, 4))
data_full_model = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10, data = data)
plot(data_full_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10)
47
43
101
6
4
2
0
−2
8
10
12
14
16


Q−Q Residuals
Standardized residuals
6
4
2
0
−2
47
43
101
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10)


Scale−Location
47
43
101
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
8
10
12
14
16
Fitted values
lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10)




```
data_null_model = lm(y ~ 1, data = data)
vif(data_full_model)
```

```
##          GVIF Df GVIF^(1/(2*Df))
## x1   1.176172  1        1.084515
## x2   2.154694  1        1.467888
## x3   1.978520  1        1.406599
## x4   1.416265  1        1.190070
## x5  35.699204  1        5.974881
## x6   1.855334  1        1.362107
## x7   1.715222  3        1.094091
## x8  34.211423  1        5.849053
## x9   7.055523  1        2.656224
## x10  3.241812  1        1.800503
```

```
x5_model = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x9 + x10, data = data)
x8_model = lm(y ~ x1 + x2 + x3 + x4 + x6 + x8 + x7 + x9 + x10, data = data)
BIC(x5_model) #remove
```

```
## [1] 427.8142
```

```
BIC(x8_model)
```

```
## [1] 417.261
```

```r
#Since the model with x5 has a higher BIC value, I would remove x5 instead of x8
#from the data and the full model
#I would also remove x9 since it also has high correlation with other predictors

data_full_model = lm(y ~ x1 + x2 + x3 + x4 + x6 + x8 + x7 + x10, data = data)
forward_model_BIC <- step(data_null_model,
                          scope = list(lower = data_null_model, upper = data_full_model),
                          direction = "forward",
                          k = log(nrow(data)),
                          trace = TRUE)
```

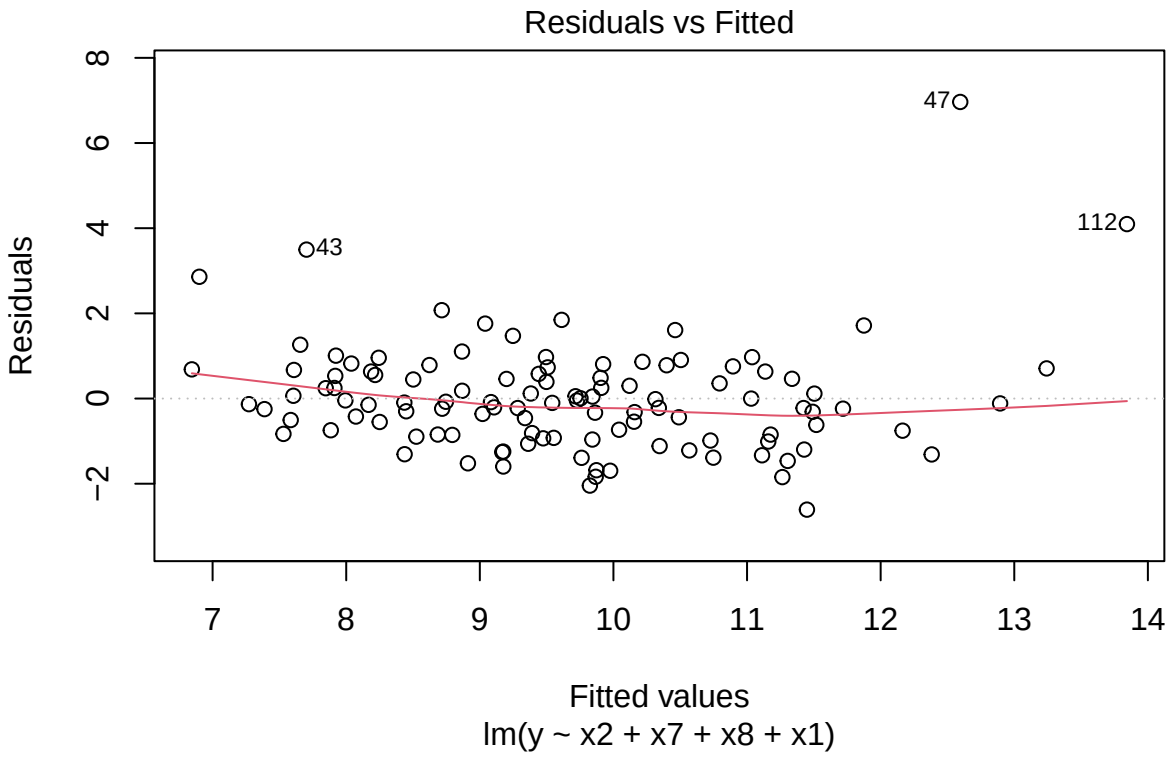
```
## Start:  AIC=150.14
## y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + x2    1   116.446 292.76 117.03
## + x8    1    91.895 317.32 126.13
## + x7    3   103.554 305.66 131.35
## + x4    1    59.864 349.35 137.00
## + x10   1    51.727 357.48 139.60
## + x3    1    43.672 365.54 142.12
## + x6    1    36.084 373.13 144.44
## <none>              409.21 150.14
## + x1    1    14.604 394.61 150.76
##
## Step:  AIC=117.03
## y ~ x2
##
##        Df Sum of Sq    RSS     AIC
## + x7    3    71.967 220.80  99.331
## + x8    1    35.020 257.75 107.360
## + x1    1    14.514 278.25 116.010
## + x6    1    12.897 279.87 116.665
## <none>              292.76 117.029
## + x4    1    10.186 282.58 117.754
## + x10   1     9.046 283.72 118.209
## + x3    1     0.480 292.28 121.570
##
## Step:  AIC=99.33
## y ~ x2 + x7
##
##        Df Sum of Sq    RSS     AIC
## + x8    1   25.6797 195.12  90.087
## + x6    1   11.4789 209.32  98.025
## + x1    1   11.1884 209.61  98.182
## <none>              220.80  99.331
## + x10   1    3.9031 216.90 102.043
## + x4    1    2.0891 218.71 102.984
## + x3    1    2.0212 218.78 103.019
##
## Step:  AIC=90.09
## y ~ x2 + x7 + x8
##
```

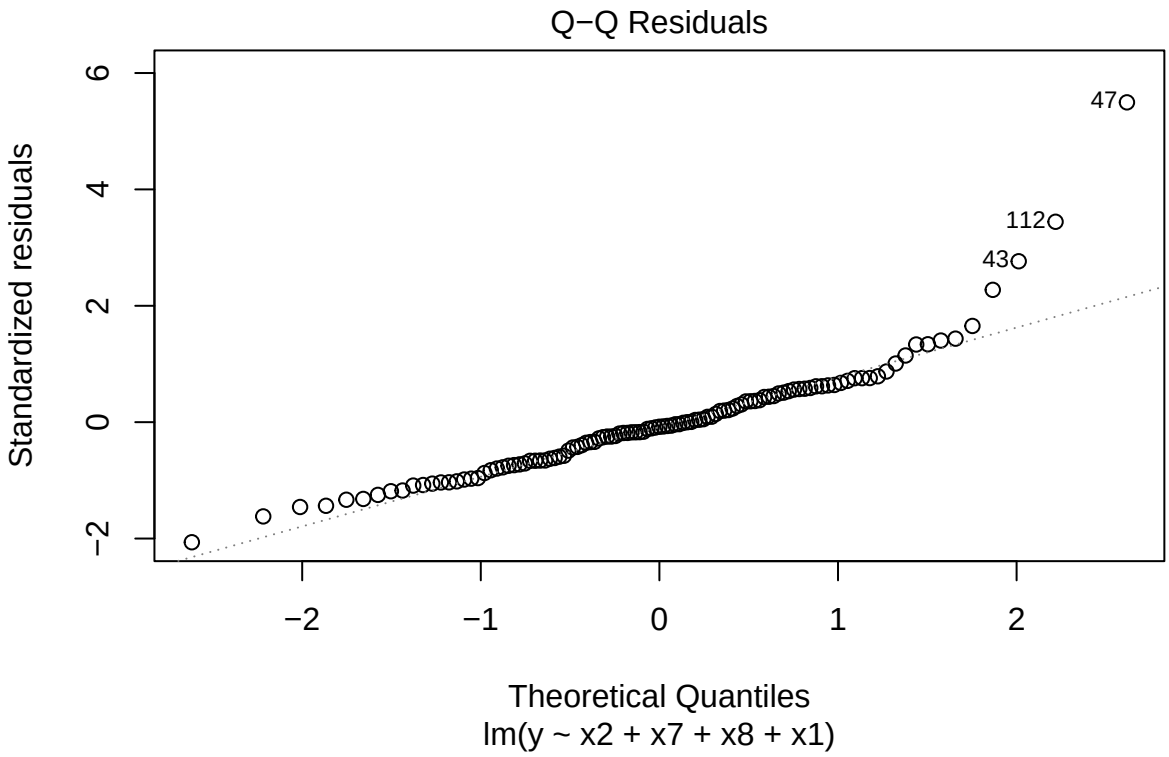
```
##        Df Sum of Sq    RSS    AIC
## + x1    1   13.0448 182.07 86.995
## <none>              195.12 90.087
## + x10   1    7.2121 187.91 90.558
## + x4    1    4.9709 190.15 91.898
## + x3    1    0.7186 194.40 94.397
## + x6    1    0.2136 194.90 94.690
##
## Step:  AIC=87
## y ~ x2 + x7 + x8 + x1
##
##        Df Sum of Sq    RSS    AIC
## <none>              182.07 86.995
## + x10   1    7.3725 174.70 87.052
## + x4    1    5.6181 176.46 88.181
## + x6    1    0.8574 181.22 91.189
## + x3    1    0.0958 181.98 91.663
```

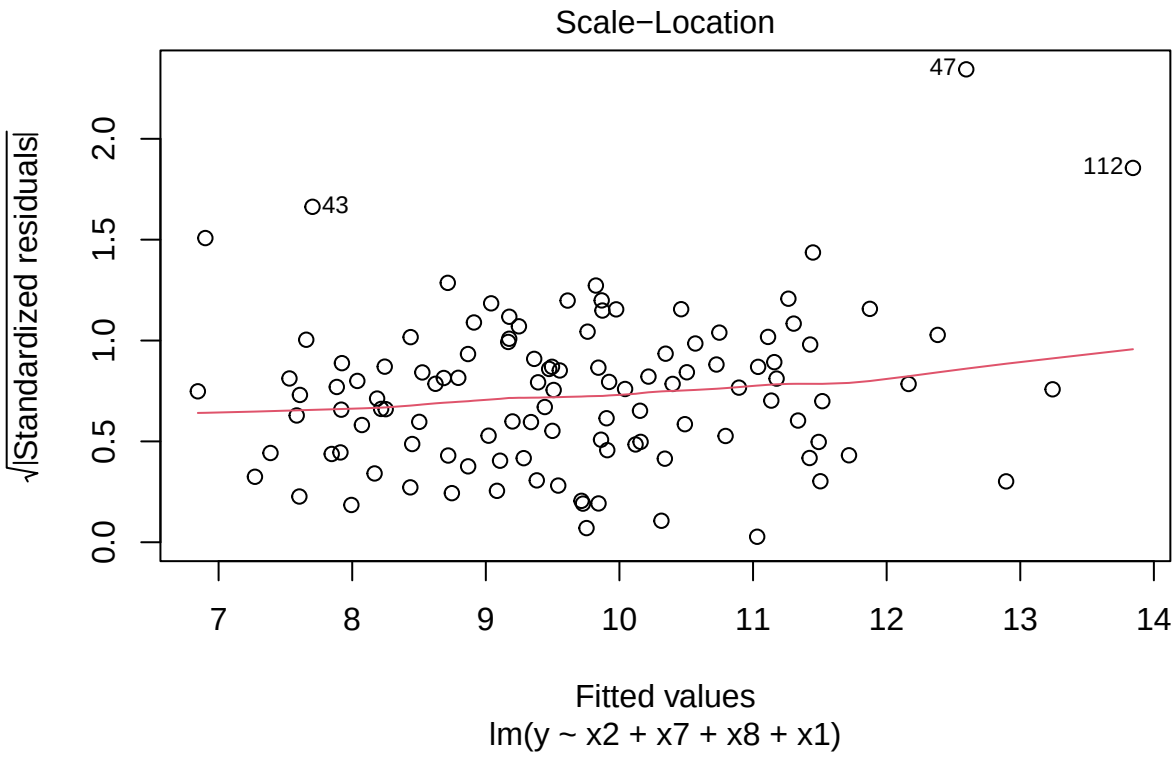
```
summary(forward_model_BIC)
```

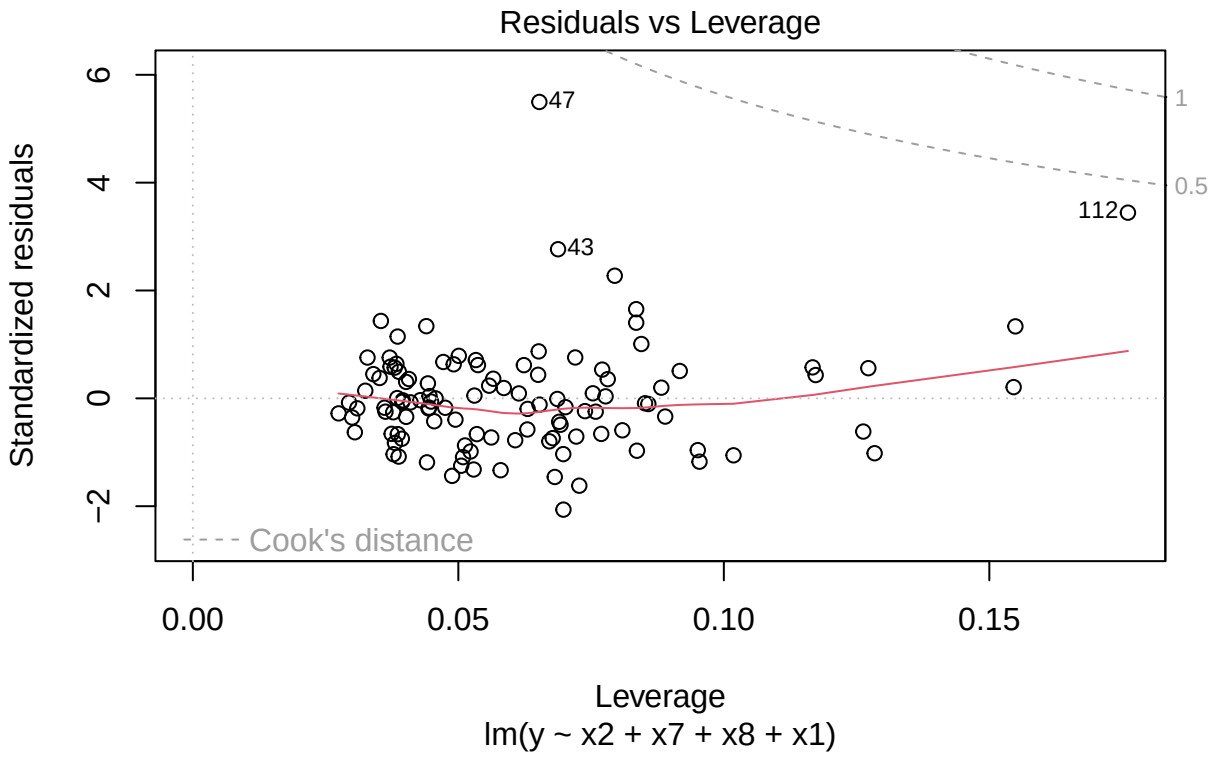
```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8 + x1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6083 -0.8454 -0.1019  0.6335  6.9643
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  3.590326   1.634615   2.196    0.030239 *
## x2           0.513546   0.103922   4.942 0.000002911 ***
## x72         -0.968717   0.349448  -2.772    0.006580 **
## x73         -1.293503   0.341274  -3.790    0.000250 ***
## x74         -2.361735   0.417885  -5.652 0.000000135 ***
## x8           0.003563   0.000890   4.004    0.000116 ***
## x1           0.078372   0.028439   2.756    0.006894 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.311 on 106 degrees of freedom
## Multiple R-squared:  0.5551, Adjusted R-squared:  0.5299
## F-statistic: 22.04 on 6 and 106 DF,  p-value: < 0.00000000000000022
```

```
plot(forward_model_BIC)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x2 + x7 + x8 + x1)
47
112
43
8
6
4
2
0
−2
7
8
9
10
11
12
13
14


Q−Q Residuals
Standardized residuals
6
4
2
0
−2
47
112
43
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x2 + x7 + x8 + x1)

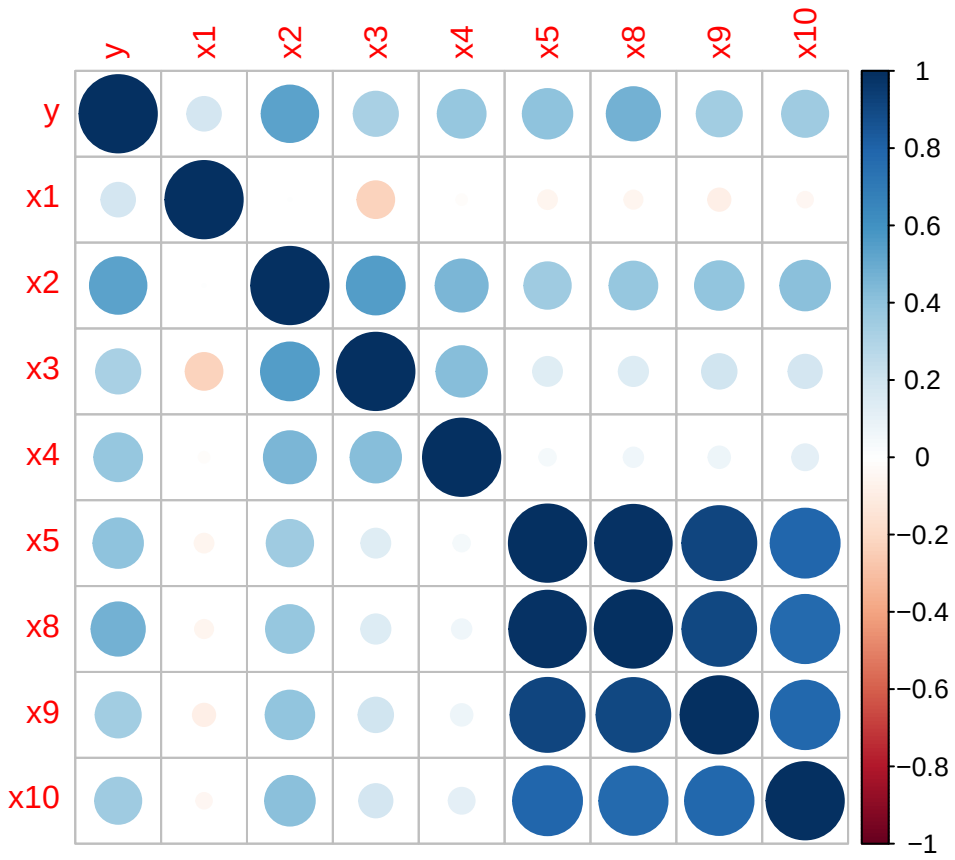

Scale−Location
47
112
43
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
7
8
9
10
11
12
13
14
Fitted values
lm(y ~ x2 + x7 + x8 + x1)




```
vif(forward_model_BIC)
```

```
##        GVIF Df GVIF^(1/(2*Df))
## x2 1.266169  1        1.125242
## x7 1.169782  3        1.026481
## x8 1.221003  1        1.104990
## x1 1.049763  1        1.024580
```

```
cor_matrix <- cor(data[, sapply(data, is.numeric)], use = "complete.obs")
corrplot(cor_matrix, method = "circle")
```



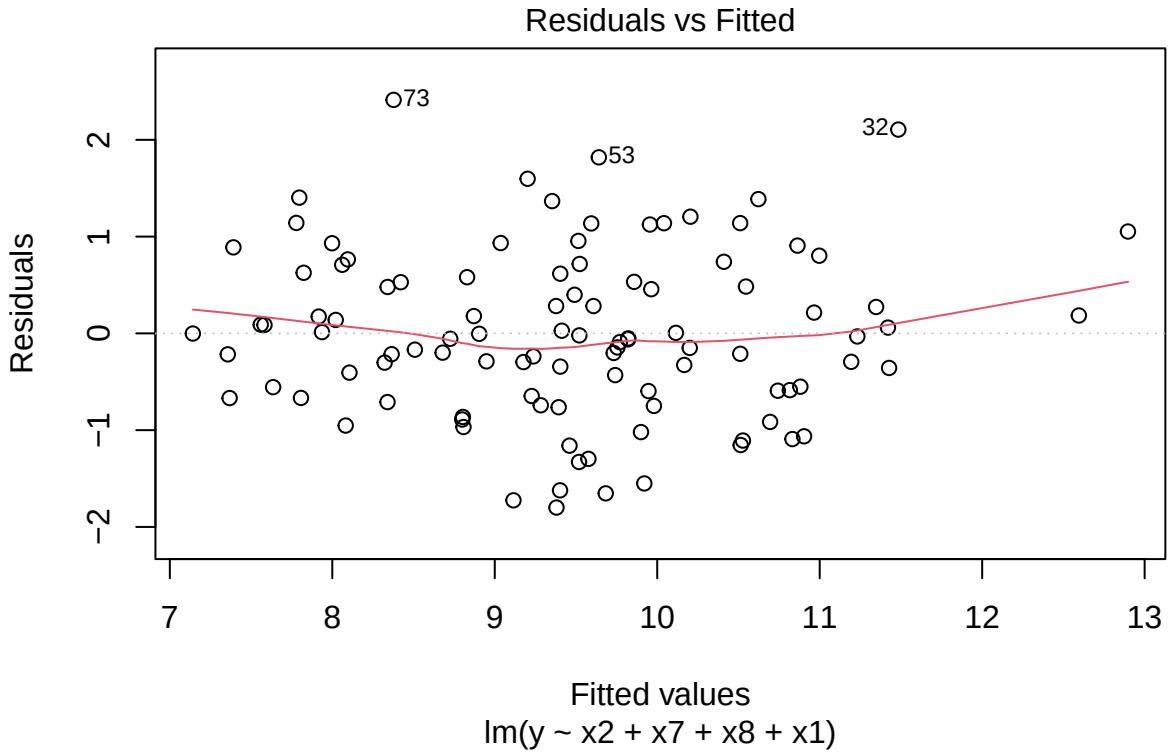

```
BIC(forward_model_BIC)
```

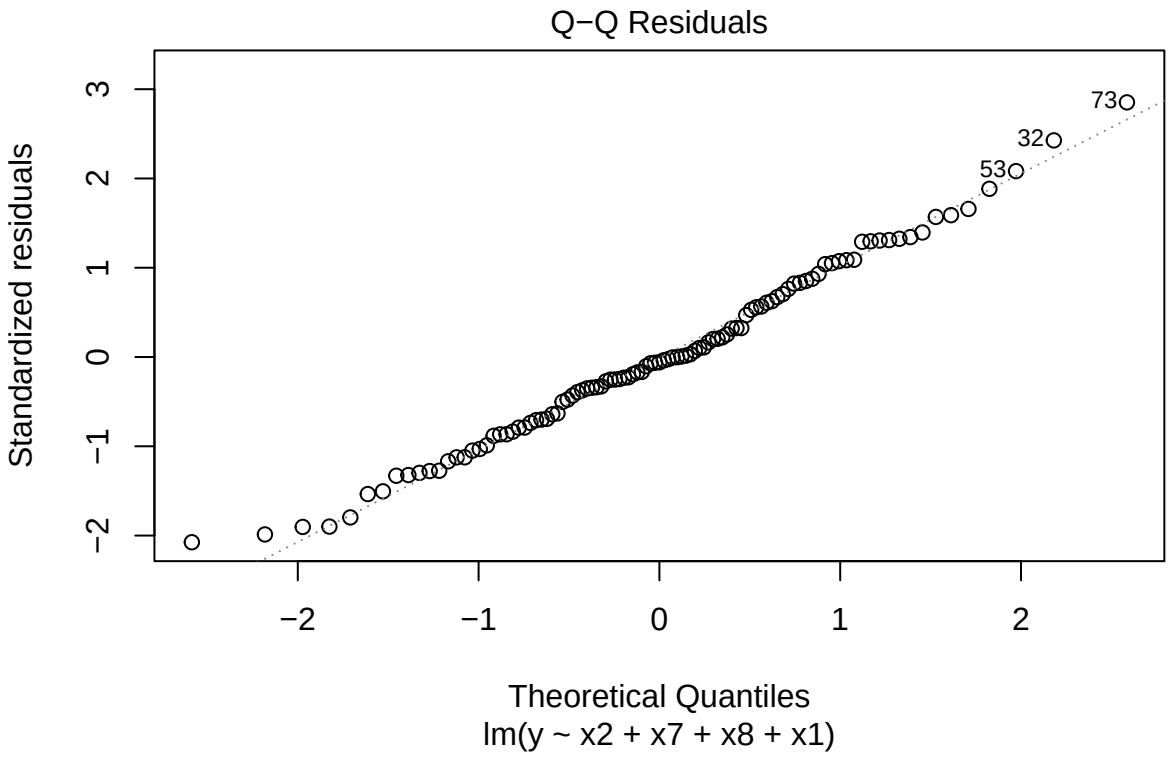
```
## [1] 412.4026
```

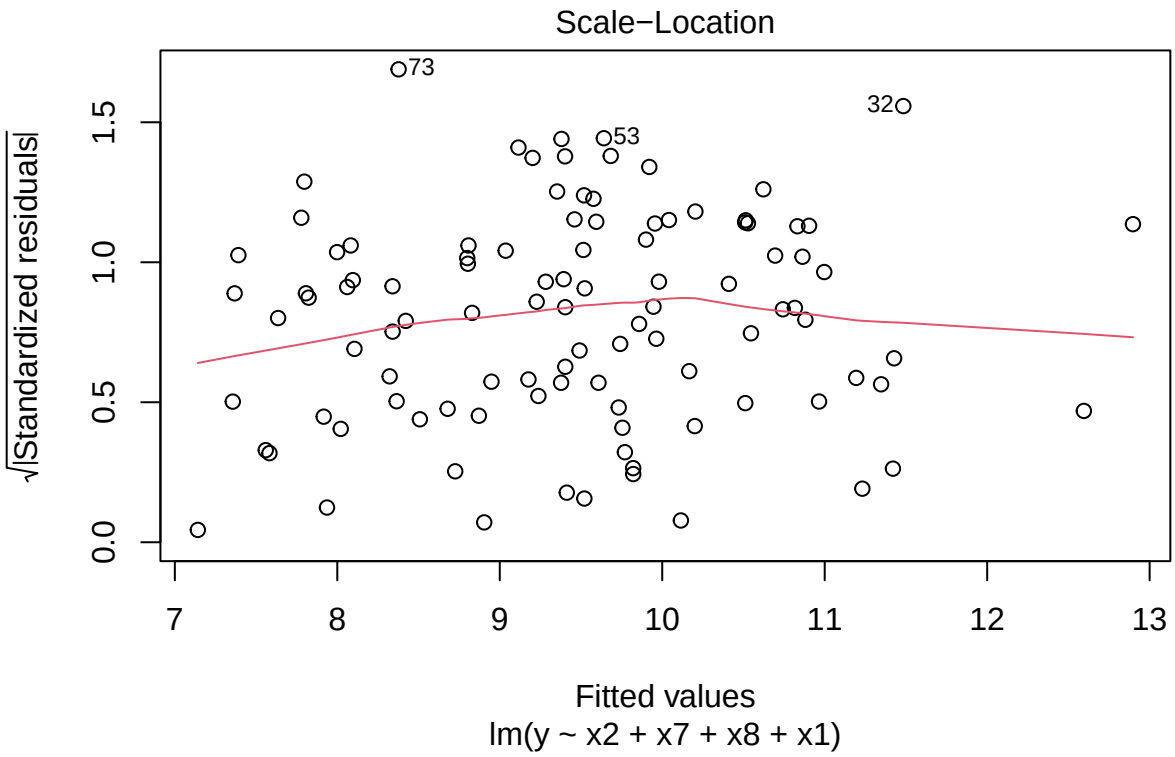
```
#Adjusted R-squared:  0.531
#Since x8 and x9 are similar and x8 better explain the variability of the data, I will
#remove x9 from the final model
final_model = forward_model_BIC
```

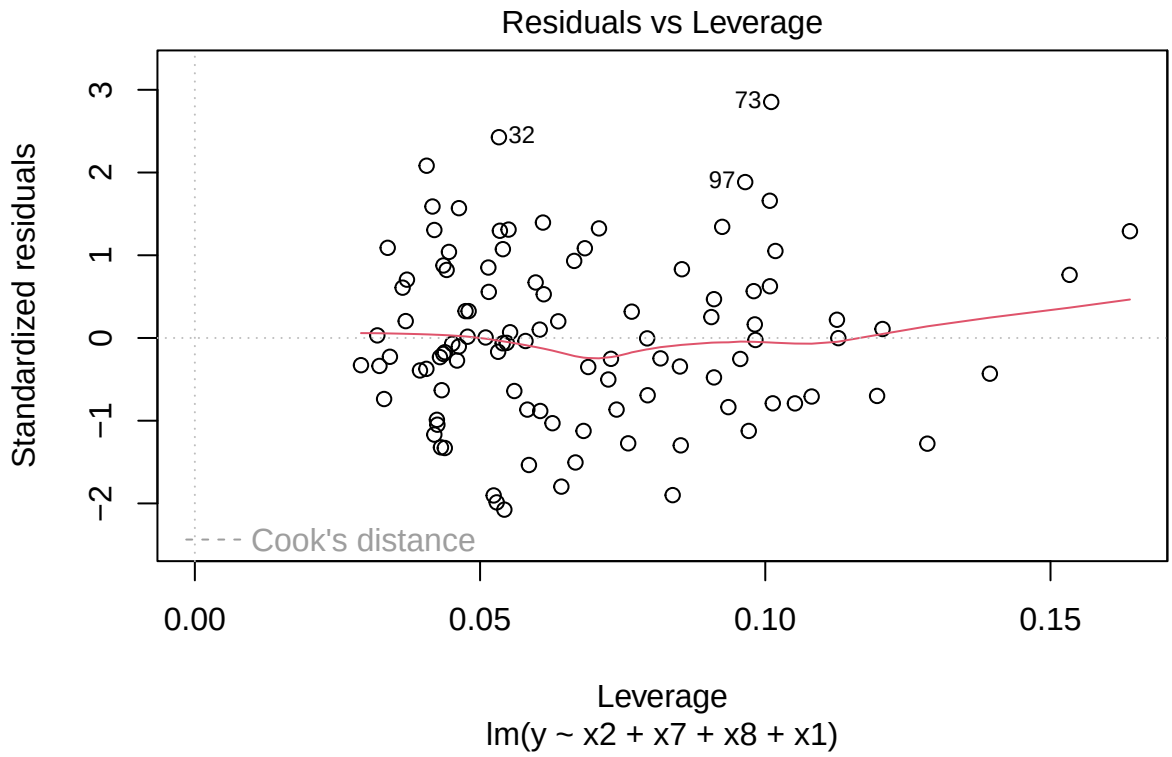
```
data$standardized_residuals = rstandard(final_model)
data$cooks_distance = cooks.distance(final_model)
data$leverage = hatvalues(final_model)

leverage_threshold = (2 * 7) / nrow(data)
data$high_leverage = data$leverage > leverage_threshold
clean_data = subset(data, leverage <= leverage_threshold)
cooks_threshold = 4 / (nrow(data) - 7 - 1)
data$high_influential = data$cooks_distance > cooks_threshold
clean_data = subset(clean_data, cooks_distance <= cooks_threshold)

# Flag and remove outliers
t_val = qt(1 - 0.05/ (2 * nrow(clean_data)), (nrow(clean_data) - 7))
data$outliers = abs(data$standardized_residuals) > t_val
clean_data = subset(clean_data, abs(standardized_residuals) <= t_val)
write.csv(data,"/Users/justin/Documents/RStudio/SENIC2_outliers.csv")
```

```r
clean_model = lm(y ~ x2 + x7 + x8 + x1, data = clean_data)
plot(clean_model)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(y ~ x2 + x7 + x8 + x1)
73
32
53
−2
−1
0
1
2
3


Scale−Location
73
32
53
√|Standardized residuals|
0.0
0.5
1.0
1.5
7
8
9
10
11
12
13
Fitted values
lm(y ~ x2 + x7 + x8 + x1)




```
summary(clean_model)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8 + x1, data = clean_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79985 -0.59434 -0.05144  0.59863  2.41275
##
## Coefficients:
##               Estimate Std. Error t value        Pr(>|t|)
## (Intercept)  2.9021489  1.2840523   2.260        0.026072 *
## x2           0.5494837  0.0812490   6.763 0.0000000010572 ***
## x72         -0.6188861  0.2466168  -2.510        0.013764 *
## x73         -1.0048801  0.2434017  -4.128 0.0000778782304 ***
## x74         -2.3477317  0.3181823  -7.379 0.0000000000573 ***
## x8           0.0020192  0.0007232   2.792        0.006321 **
## x1           0.0872021  0.0222501   3.919        0.000167 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.892 on 96 degrees of freedom
## Multiple R-squared:  0.6506, Adjusted R-squared:  0.6288
## F-statistic:  29.8 on 6 and 96 DF,  p-value: < 0.00000000000000022
```

```r
#Shapiro-Wilks Test for Normality
shapiro.test(resid(clean_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(clean_model)
## W = 0.99045, p-value = 0.6818
```

```r
#Shapiro p-value = 0.6818
#Since p-value is larger than alpha = 0.05, I fail to reject the null hypothesis
#Normality assumption satisfied

#Fligner Killeen Test for constant variance
fitted_value = fitted(clean_model)
fitted_groups <- ifelse(fitted_value <= median(fitted_value), "Group 1", "Group 2")
fligner.test(resid(clean_model) ~ fitted_groups)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  resid(clean_model) by fitted_groups
## Fligner-Killeen:med chi-squared = 0.39993, df = 1, p-value = 0.5271
```

```r
#Fligner p-value = 0.5271
#Since p-value is greater than alpha = 0.05, I fail to reject the null hypothesis and
#conclude that there is equal variance across two groups
#Equal variance assumption satisfied

#Levene Test for constant variance
levene_data = data.frame(x = as.factor(fitted_groups), y = clean_data$y)
levene_result = leveneTest(y ~ x, levene_data)
levene_result
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  2.1221 0.1483
##       101
```

```r
#Levene p-value = 0.1483
#Since p-value is greater alpha = 0.05, I fail to reject the null hypothesis and
#conclude that there is equal variance across the two groups
```

```r
#Simultaneous Confidence Intervals
confint(clean_model, level = 1 -0.05 / length(coef(clean_model)))
```

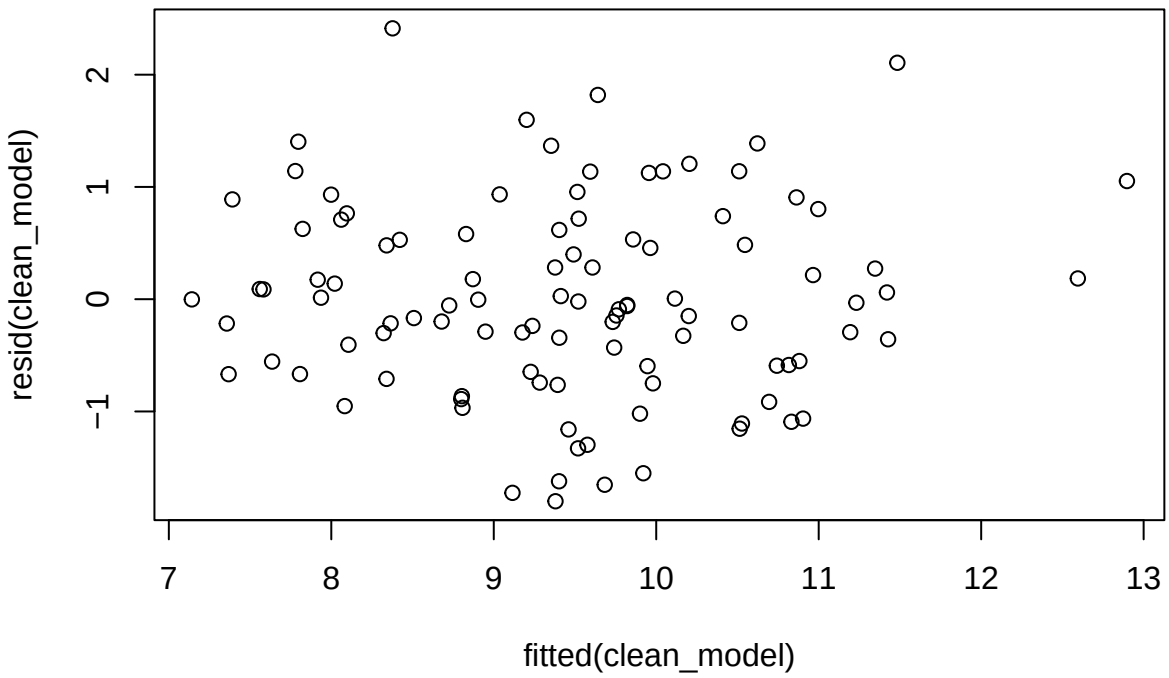
```
##                  0.357 %     99.643 %
## (Intercept) -0.627694025  6.431991874
## x2           0.326131293  0.772836053
## x72         -1.296832563  0.059060291
## x73         -1.673988153 -0.335772103
```

```
## x74         -3.222410709 -1.473052775
## x8           0.000031139  0.004007313
## x1           0.026036850  0.148367372
```

```
#                   0.357 %     99.643 %
# (Intercept) -0.627694025  6.431991874
# x2           0.326131293  0.772836053
# x72         -1.296832563  0.059060291
# x73         -1.673988153 -0.335772103
# x74         -3.222410709 -1.473052775
# x8           0.000031139  0.004007313
# x1           0.026036850  0.148367372


#Ho: Beta_i = 0, i = 1, 2, 7, 8. The predictor does not contribute significantly to the model
#Ha: Beta_i != 0, i = 1, 2, 7, 8. The predictor contributes significantly to the model

anova(clean_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value              Pr(>F)
## x2         1 72.697  72.697 91.3701 0.00000000000000132 ***
## x7         3 52.989  17.663 22.1998 0.00000000005298711 ***
## x8         1  4.336   4.336  5.4493           0.0216607 *
## x1         1 12.221  12.221 15.3599           0.0001666 ***
## Residuals 96 76.381   0.796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Analysis of Variance Table
#
# Response: y
#           Df Sum Sq Mean Sq F value              Pr(>F)
# x2         1 72.697  72.697 91.3701 0.00000000000000132 ***
# x7         3 52.989  17.663 22.1998 0.00000000005298711 ***
# x8         1  4.336   4.336  5.4493           0.0216607 *
# x1         1 12.221  12.221 15.3599           0.0001666 ***
# Residuals 96 76.381   0.796
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

plot(fitted(clean_model), resid(clean_model))
```




```
#Residuals has no apparent patterns and scattered evenly across the plot suggesting linearity.
```