Report on SENIC Data Set

Saikat Bera, Presidency University, KOLKATA

December 1, 2022

Abstract

The **SENIC** data set contains information on certain variables from which we are interested in using age, number of beds, infection risk, and available facilities/services to predict the average length of hospital stay of all patients in hospital.

We first load the dataset and check the head so as to confirm our load:

```
##   id  stay  age infectionRisk cultRatio chestXrayRatio numBeds medSchool region
## 1  1  7.13 55.7           4.1       9.0           39.6     279         2      4
## 2  2  8.82 58.2           1.6       3.8           51.7      80         2      2
## 3  3  8.34 56.9           2.7       8.1           74.0     107         2      3
## 4  4  8.95 53.7           5.6      18.9          122.8     147         2      4
## 5  5 11.20 56.5           5.7      34.5           88.9     180         2      1
## 6  6  9.76 50.9           5.1      21.9           97.0     150         2      2
##   avgDailyCensus numNurses facilities
## 1            207       241         60
## 2             51        52         40
## 3             82        54         20
## 4             53       148         40
## 5            134       151         40
## 6            147       106         40
```

The 12 variables are:

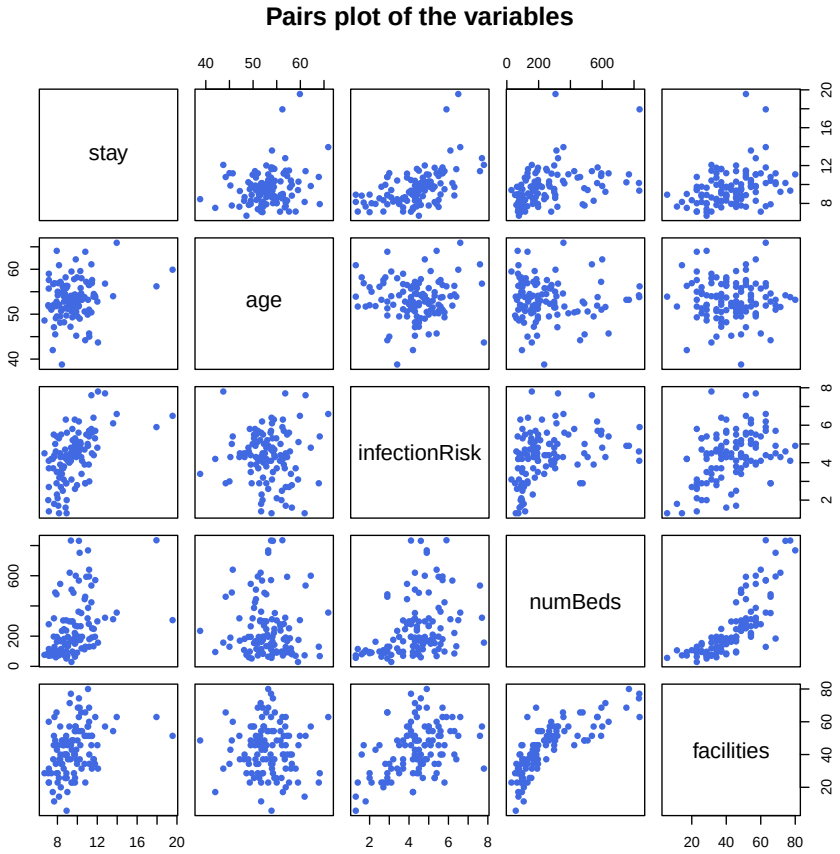
- id: Identification number of hospital, 1-113
- stay: Average length of stay of all patients in hospital (in days)
- age: Average age of patients (years)
- infectionRisk: Average estimated probability of acquiring infection in hospital (in percentage)
- cultRatio: Routine culturing ratio, which is the ratio of number of cultures performed to number of patient without signs or symptoms of hospital acquired infection, times 100.

- chestXrayRatio: Routine chest x-ray ratio, which is the ratio of number of x-rays performed to the number of patients without signs or symptoms of pneumonia, times 100.
- numBeds: Average number of beds in the hospital during study period.
- medSchool: Medical school affiliation
  - 1 = yes
  - 2 = no
- region: Geographic region
  - 1 = NE
  - 2 = NC
  - 3 = S
  - 4 = W
- avgDailyCensus: Average daily census, which is the average number of patients in hospital per day during study period.
- numNurses: Average number of full time equivalent registered and licensed practical nurses during study period (number full time + one half the number part time).
- facilities: Available facilities and services, which is percent of 35 potential facilities and services that are provided by the hospital.

We extract our required columns and we'll use this modified data for our work further:

```
##    stay  age infectionRisk numBeds facilities
## 1  7.13 55.7           4.1     279         60
## 2  8.82 58.2           1.6      80         40
## 3  8.34 56.9           2.7     107         20
## 4  8.95 53.7           5.6     147         40
## 5 11.20 56.5           5.7     180         40
## 6  9.76 50.9           5.1     150         40
```

We create a pairs plot from the correlaton matrix:



We cannot conclude any specific relationships from this pairs plot. However, the number of beds and facilities have a somewhat linear relatioship.

We might want to fit a linear model to the relationship between stay and facilities:

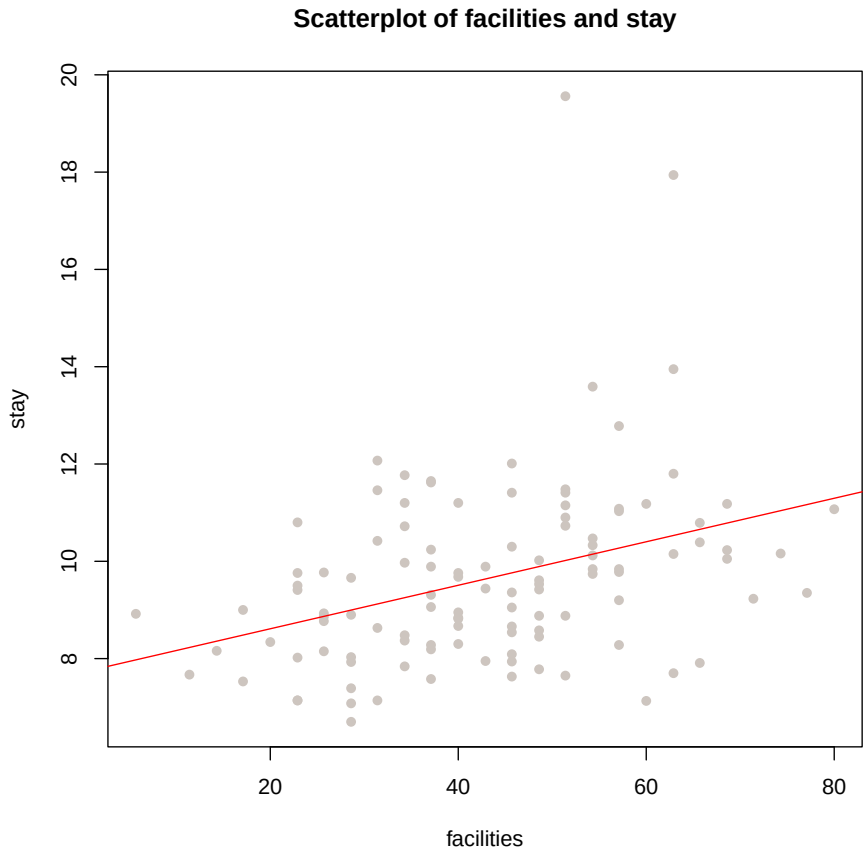
```
##
## Call:
## lm(formula = stay ~ facilities, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2712 -1.0716 -0.2816  0.7584  9.5433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.71877    0.51020  15.129  < 2e-16 ***
```

```
## facilities   0.04471    0.01116   4.008 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.795 on 111 degrees of freedom
## Multiple R-squared:  0.1264,Adjusted R-squared:  0.1185
## F-statistic: 16.06 on 1 and 111 DF,  p-value: 0.0001113
```

The estimated slope and its standard error are as follows:

```
##   Estimate Std. Error
## 0.04470767 0.01115550
```

We then continue and plot a graph of the two variables and our estimated linear model:



We move on to finding the linear model for the stay variable with the regressors being all the 4 other variables:

```
##
## Call:
## lm(formula = stay ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8934 -0.9221 -0.1140  0.7819  7.8620
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.813602   1.828061   0.992  0.32337
## age            0.087470   0.032468   2.694  0.00819 **
## infectionRisk  0.640635   0.118581   5.403 3.95e-07 ***
## numBeds        0.003169   0.001238   2.560  0.01186 *
## facilities    -0.009508   0.016069  -0.592  0.55528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.53 on 108 degrees of freedom
## Multiple R-squared:  0.3823,Adjusted R-squared:  0.3594
## F-statistic: 16.71 on 4 and 108 DF,  p-value: 1.094e-10
```

The fitted model is:

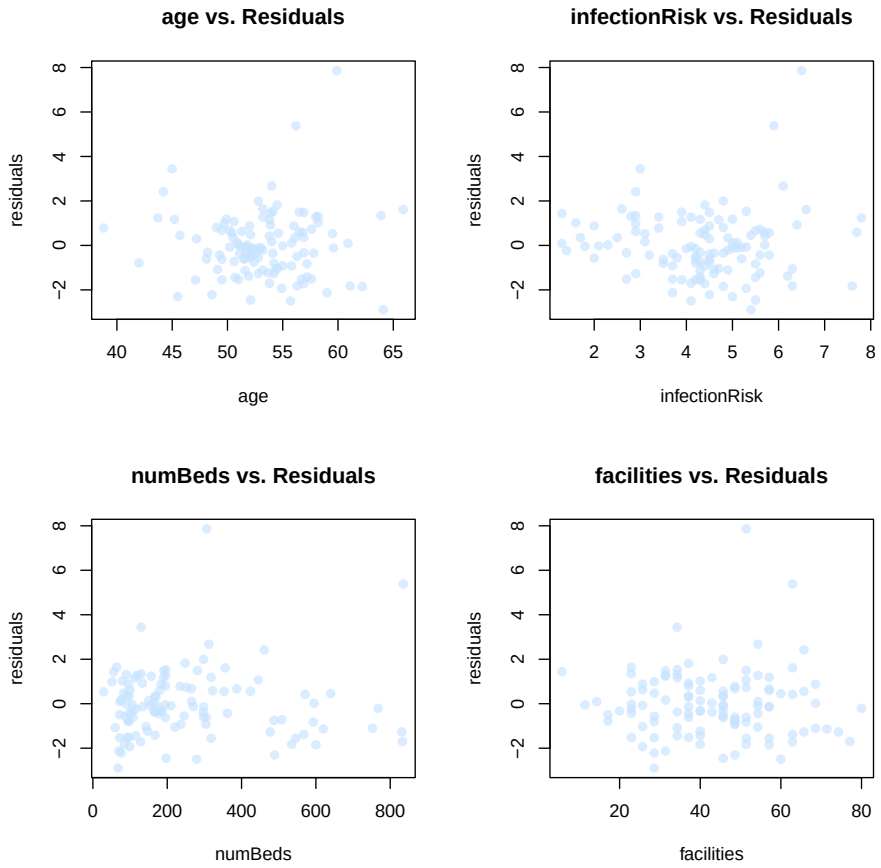
$$y = 1.813602 + 0.087470x_1 + 0.640635x_2 + 0.003169x_3 - 0.009508x_4$$

where:

- y: the average length of hospital stay of all patients in hospital
- x_1: the average age of patients (in years)
- x_2: the average estimated probability of acquiring infection in hospital (in percentage)
- x_3: the average number of beds in the hospital during study period
- x_4: the available facilities and services

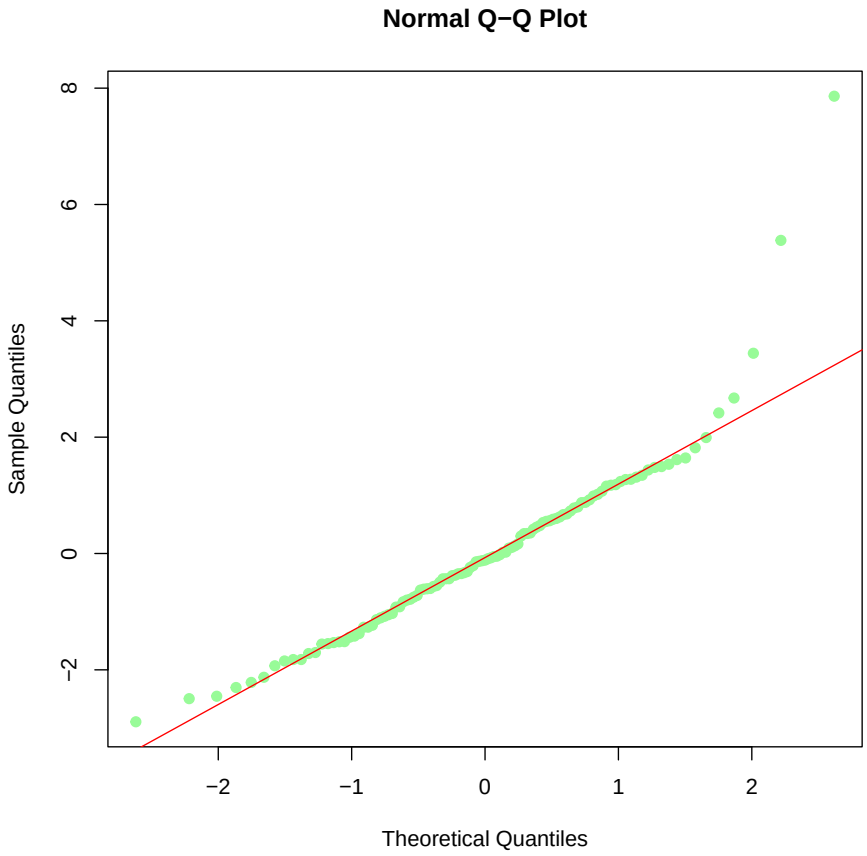
The coefficients on service differs in the two types of linear regression. We note that in the first case, the p-value is 0.000111 which rejects the null hypothesis at 5% level that the coeficient is not statistically significant. Hence, in the first case, we accept the alternative that it is statistically significant. However, in the second case, the p-value comes out as 0.55528. Thus, we accept the null hypothesis at 5% level that the coefficient is statistically significant.

We plot the residual vs predictor variable graphs for all the predictor variables:




```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
```

We plot the qqplot and find the following graph:



We note that, while keeping the other variables fixed, with every unit change in the average measurement of they stay variable, the value of infection risk will change by 0.640635 on an average. We find the p-value of our test to be:

```
## [1] 3.950788e-07
```

Thus, we reject our null hypothesis that the slope is insignificant with $\alpha = 0.05$ level of significance.

After fitting the model, we find the root mean squared error to be:

```
## Warning:  package 'Metrics' was built under R version 4.2.2

## [1] 1.495684
```

We note that the root mean squared error is very low. Hence, we can conclude that the model in general is a good fit.

We obtain an interval estimate of the expected value of average length of hospital stay when average age=54, number of beds=100, infection risk=5%, and service=30% at 95% confidence interval and we get it as:

```
##        fit      lwr      upr
## 1 9.771774 9.305479 10.23807
```

This implies, that if we take 100 samples with the given conditions, in 95 times, our predicted variable will lie in between these two limits.

We obtain a prediction interval for the average length of stay for a new hospital with average age=58, number of beds=200, infection risk=6%, and service=40% at 99% confidence interval and we get it as:

```
##        fit      lwr      upr
## 1 10.98406 6.897514 15.07061
```