Linear Regression for diabetes patients re-admitted in hospital within 30 days

Neeraj Kumar Reddy Panta
Ruchi Dilip Kukde

Linear Regression

- Focus : Patients readmitted in hospital within 30 days.
- Data used for regression is numeric
- Units of Variables:
  - time_in_hospital: Days
  - num_medications : Count
  - num_lab_procedures: Count

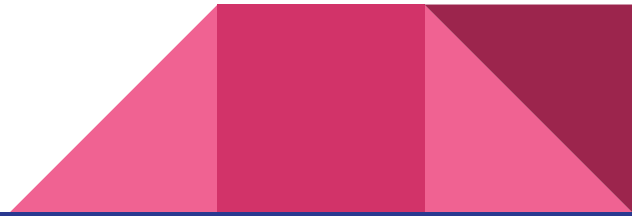
time_in_hospital	num_medications	num_lab_procedures
1	1	41
3	18	59
2	13	11
2	16	44
1	8	51
3	16	31
4	21	70
5	12	73
13	28	68
12	18	33
9	17	47
7	11	62
7	15	60
10	31	55
1	2	49
12	13	75
4	17	45
3	11	29
5	23	35
6	23	42

Simple Linear Regression Model Overview

For simple Linear Regression model we want to see how the value of time_in_hospital(in days) varies for patients who have been re-admitted in hospital within 30 days based on the number of medications administered

- Outcome - time_in_hospital $\hat{Y}$
- Predictor - Num_medications X_1

- Equation of the model: $\hat{Y} = b_1 X_1 + b_0$



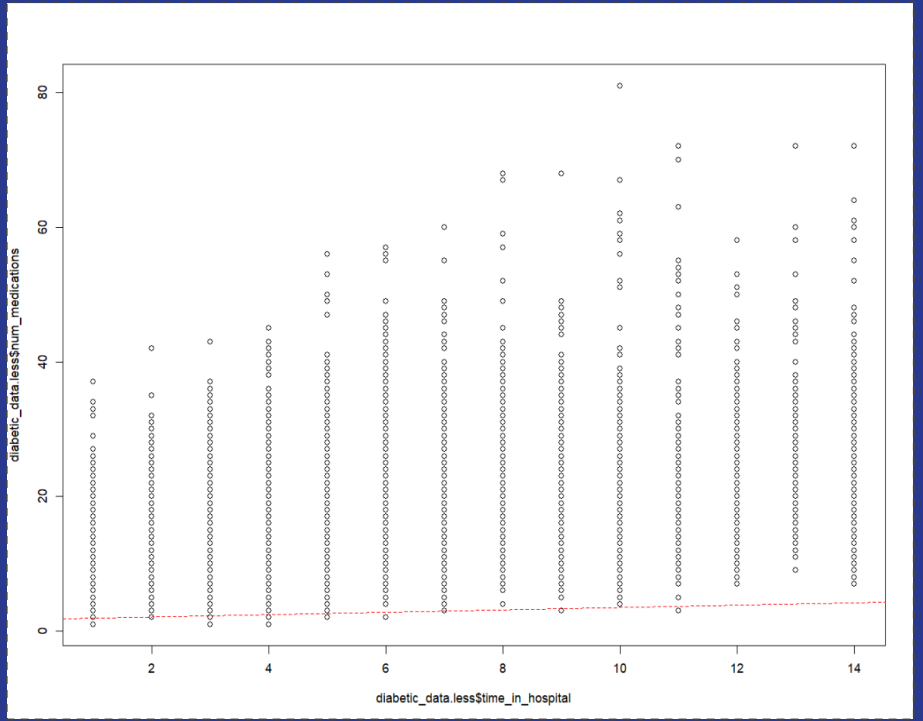
Preliminary Assessment for Simple Linear Regression Model

Verifying Linearity with Scatter Plot

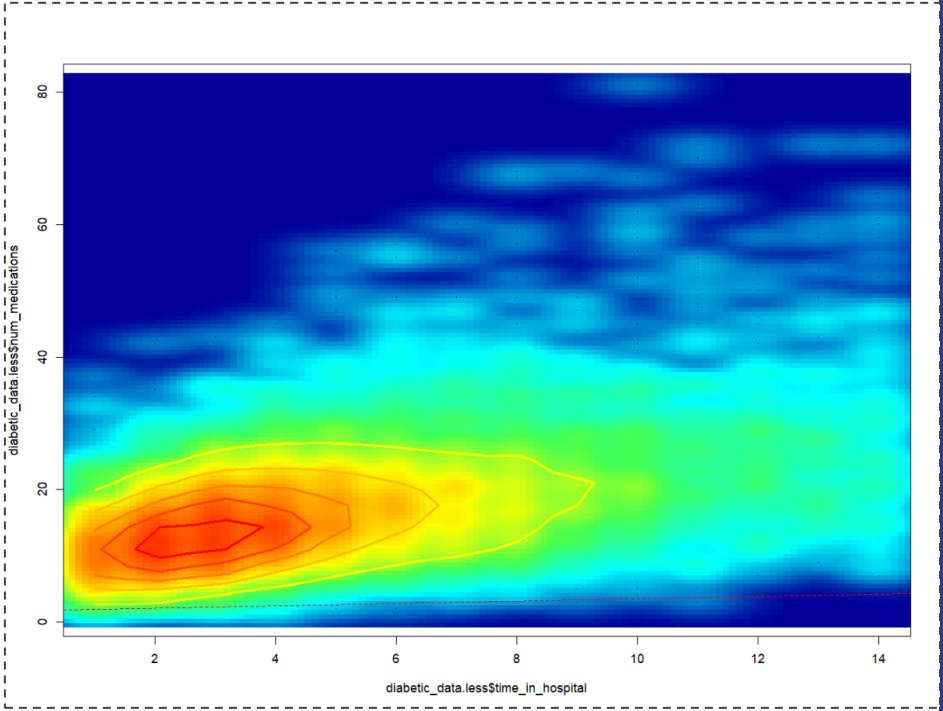
Observation:

We cannot depict the relationship between the two variables time_in_hospital and num_medications as the variables are measured on a discrete scale, this presents us with a challenge.

However, there could still be a linear relationship which is difficult to observe using this display



Verifying Linearity with Heat Map R Scatter Plot



We generate this graph using Heat Mapping along-with base R functions such as *Smoothscatter* and *contour* on a scatter plot

Looks like a positive correlation!

But it is better to compute the Pearson and Spearman correlation coefficients and then assess the strength

Pearson & Spearman correlation coefficient tests

Both Pearson and Spearman correlation coefficient tests for patients readmitted in hospital within 30 days show that there is a ***positive correlation*** between the variables time_in_hospital and num_medications.

```
> cor.test(x = diabetic_data.less$time_in_hospital, y=diabetic_data.less$num_medications,
+          alternative = "two.sided",method = "pearson")

        Pearson's product-moment correlation

data:  diabetic_data.less$time_in_hospital and diabetic_data.less$num_medications
t = 58.207, df = 11355, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4650899 0.4934226
sample estimates:
      cor
0.4793812
```

```
> cor.test(x = diabetic_data.less$time_in_hospital, y=diabetic_data.less$num_medications,
+          alternative = "two.sided",method = "spearman")

        Spearman's rank correlation rho

data:  diabetic_data.less$time_in_hospital and diabetic_data.less$num_medications
S = 1.286e+11, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4732401
```

Output of Simple Regression Model

time_in_hospital = 0.17929 (num_medications) + 1.73771

```
> summary(diabetic_less_M.1)

Call:
lm(formula = time_in_hospital ~ num_medications, data = diabetic_data.less)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3714 -1.8892 -0.5306  1.4215 11.0073

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.73771    0.05773   30.10   <2e-16 ***
num_medications  0.17929    0.00308   58.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.658 on 11355 degrees of freedom
Multiple R-squared:  0.2298,    Adjusted R-squared:  0.2297
F-statistic:  3388 on 1 and 11355 DF,  p-value: < 2.2e-16
```

Interpreting the Output

The num_medications coefficient suggests that for every 10 count increase in the count of medications of the patient, we can expect increase in length of stay by 0.17929 * 10 = 1.7929 days, on average.

Is the Slope statistically significant?

- Yes. It is because the value of slope (b_1) has to be significantly different from 0, and in our case (0.17929 - 0) is different from zero.
- Observation from 95% confidence interval test, we can say with 95% confidence that slope lies between 0.1732 and 0.1853.
- Also, if the coefficient is large compared to the standard error, then statistically our coefficient will not be zero.

```
> confint(diabetic_less_M.1, level = 0.95)
                    2.5 %    97.5 %
(Intercept)     1.6245487 1.8508677
num_medications 0.1732509 0.1853263
```

Hypothesis Test for the slope Coefficient

$$H0: \beta = 0$$

$$H1: \beta \neq 0$$

H0 : No Linear relationship between time_in_hospital and num_medications

H1: There is Linear relationship between time_in_hospital and num_medications

- Summary shows that num_medications coefficient is 58.21 standard errors away from zero and it is pretty far from zero. The larger our t-statistic is the more certain we can be that the coefficient in not zero.

- The p-value is calculated is using t-statistic from the t distribution and it also helps us understand how significant is our coefficient is to the model. In practical terms any p-value below 0.05 is significant. In our model, we can see that *Intercept* and *num_medications* have p-value of 2e-16 which is extremely small and it is even below 0.001. We can conclude that the coefficients in this model are not zero.

- The residual Standard error is a measure of how well the model fits the data. From our model summary, we can see that on average, the actual values are 2.658 Days away from regression line.

- The F-statistic and overall p-value help us determine the result of our Hypothesis test. It is common for the F-statistic to be close to 1 if we have lot of predictors. However for smaller models, a larger F-statistic and a small p-value generally indicates that null hypothesis should be rejected and it clearly indicates that the coefficient in the model isn't zero.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.73771    0.05773   30.10   <2e-16 ***
num_medications  0.17929    0.00308   58.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.658 on 11355 degrees of freedom
Multiple R-squared:  0.2298,    Adjusted R-squared:  0.2297
F-statistic:  3388 on 1 and 11355 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression Model Overview

For Multiple Linear Regression model we want to see how the value of time_in_hospital(in Days) varies for patients who have been re-admitted in hospital within 30 days based on the Number of Medications they have used also Number of lab procedures they have undergone.

- Outcome - time_in_hospital ($\hat{Y}$)
- Predictors - Num_medications (X_1), Num_lab_procedures (X_2)

- Equation of the model: $\hat{Y} = b_1 X_1 + b_2 X_2 + b_0$

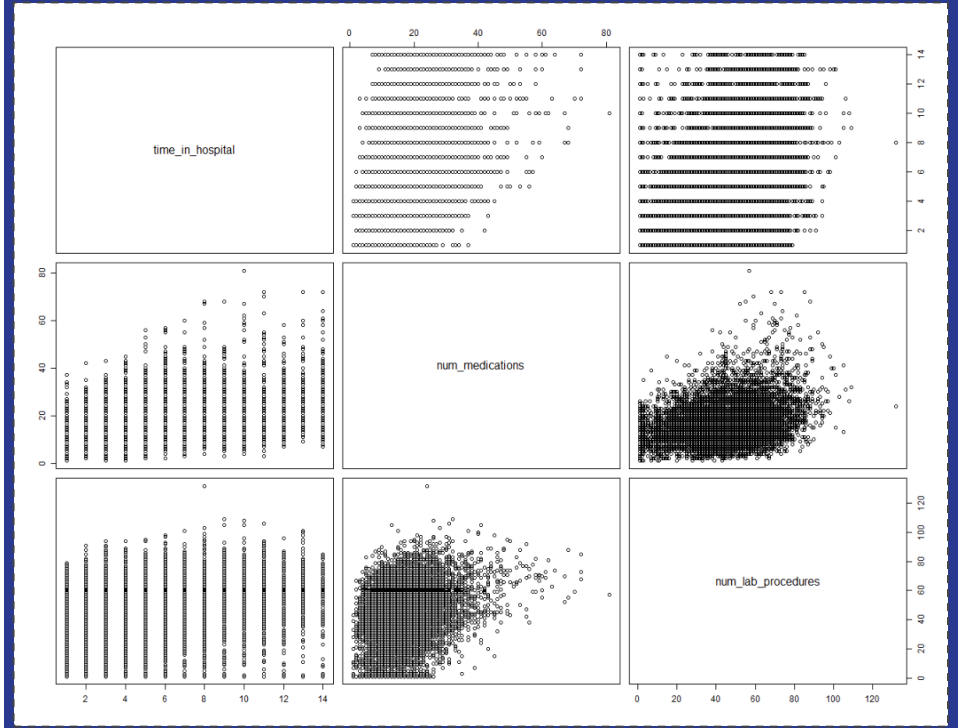
Preliminary Assessment for Multiple Linear Regression Model

Verifying Linearity with Scatter Plot

We already know from simple linear regression that time_in_hospital has a moderate positive correlation with num_procedures, but we don't know what is the relation between the other variables.

With function (*cor*) we can identify correlation between multiple variables,

```
> cor(diabetic_multi)
                   time_in_hospital num_medications num_lab_procedures
time_in_hospital          1.0000000       0.4793812          0.3176553
num_medications           0.4793812       1.0000000          0.3009453
num_lab_procedures        0.3176553       0.3009453          1.0000000
```



Output of Multiple Linear Regression Model

time_in_hospital = 0.1578 (Num_medications) + 0.0299 (Num_lab_procedures) + 0.7758

```
> summary(diabetic_less_M.2)

Call:
lm(formula = time_in_hospital ~ num_medications + num_lab_procedures,
    data = diabetic_multi)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9334 -1.8217 -0.5187  1.3418 11.9016

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.775825   0.070760   10.96   <2e-16 ***
num_medications   0.157830   0.003160   49.95   <2e-16 ***
num_lab_procedures 0.029951   0.001327   22.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.6 on 11354 degrees of freedom
Multiple R-squared:  0.2629,    Adjusted R-squared:  0.2627
F-statistic:  2024 on 2 and 11354 DF,  p-value: < 2.2e-16
```

Interpreting the output

In this case the p-value of F-statistic is <2.2e-16, which is highly significant, this means at least one of the predictor variables is significantly related to the outcome variable.

It can be seen that change in number of num_medications and num_lab_procedures, the time_in_hospital of a patient is associated.

For instance, for 10 count increase in the number of medications taken by the patient, we can expect an increase of 0.1578 * 10 = 1.578 days of patient staying in hospital (when the num_lab_procedures are constant)

Confidence interval of the model coefficients can be extracted as follows:

```
> confint(diabetic_less_M.2)
                      2.5 %     97.5 %
(Intercept)       0.57467654 0.66329459
num_medications   0.14865545 0.15269200
num_lab_procedures 0.03079195 0.03245947
```

Goodness of Fit?

- The adjusted R-squared value for multiple regression model is 0.2627, meaning that 26.27% of the variance in measure of days can be predicted by num_medications and num_lab_procedures number count.
- This model is better than the simple linear model with only num_medications which had an adjusted R-squared of 0.2297.
- The RSE gives us a measure of error of prediction. Multiple linear regression model gives an error rate of 54% , which is better than the simple linear regression model where the RSE was 0.558 (i.e. 55.8% error rate).

Thank You