Residual Analysis and Test for Assumption of normality of residuals

Srikar

2021-09-23

Introduction:

Simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns with one independent variable and one dependent variable .

Residual is the difference between an observed value of the response variable and the value of the response variable predicted from the regression line.

Objective:

1) To predict the marks based on the number of hours studied
2) To plot the residuals and test the assumptions of normality of the residuals

Data Description:

The given dataset describes the number of hours a student studied and the mask scored by them. The dataset contains two variables i.e. Hours and Scores. It contains the data of 25 students.

```
library(readxl)
dat <- read_excel("C:/Users/Srikar/Desktop/SS/R/Sem 5/Linear
Regression/Practical 8/dataset.xlsx")
head(dat)
```

	Table 1:Dataset	
	Hours	Scores
1	2.5	21
2	5.1	47
3	3.2	27
4	8.5	75
5	3.5	30
6	1.5	20

Procedure

1) Constructing the regression model

```
mod=lm(dat$Scores~.,data=dat)
summary(mod)
```

Residual Table			
Min 1Q	Median	3Q	Max
-15.918	1.839	4.593	7.265

Regression Summary Table			
Estimate	**Std. Error**	**t value**	**P-Value**
(Intercept)	2.5317	0.981	0.337
Hours	0.4529	21.583	<2e-16 ***

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.603 on 23 degrees of freedom
## Multiple R-squared:  0.9529, Adjusted R-squared:  0.9509
## F-statistic: 465.8 on 1 and 23 DF,  p-value: < 2.2e-16
```

We obtain the model :

Y = 2.4837+ 9.7758*X where

 # Y is the number of marks obtained and X is the number of hours studied by student.

We observe that the intecept p-value is above the significance value (0.05) and hence we can say that the intecept is not significant in the prediction. This means that if the values of the regressors were all zero, the intercept would not predict the average score i.e. if the student did not study at all.

The p-values of the regressor Hours shows significance as its below 0.05. That means these variables describe the linear relationship with the independent variable. The overall p-value is also lesser than significane level (0.05) and hence we can say that the model is significant.

The R-squared value 0.9529 which means that 95.29% of the variation is explained by

the regressors.

#2) Plotting the regression line and its residuals

```
library(broom)
```

```
model_diagnostics=augment(mod)
model_diagnostics
```

Table 2:Observed vs Expected

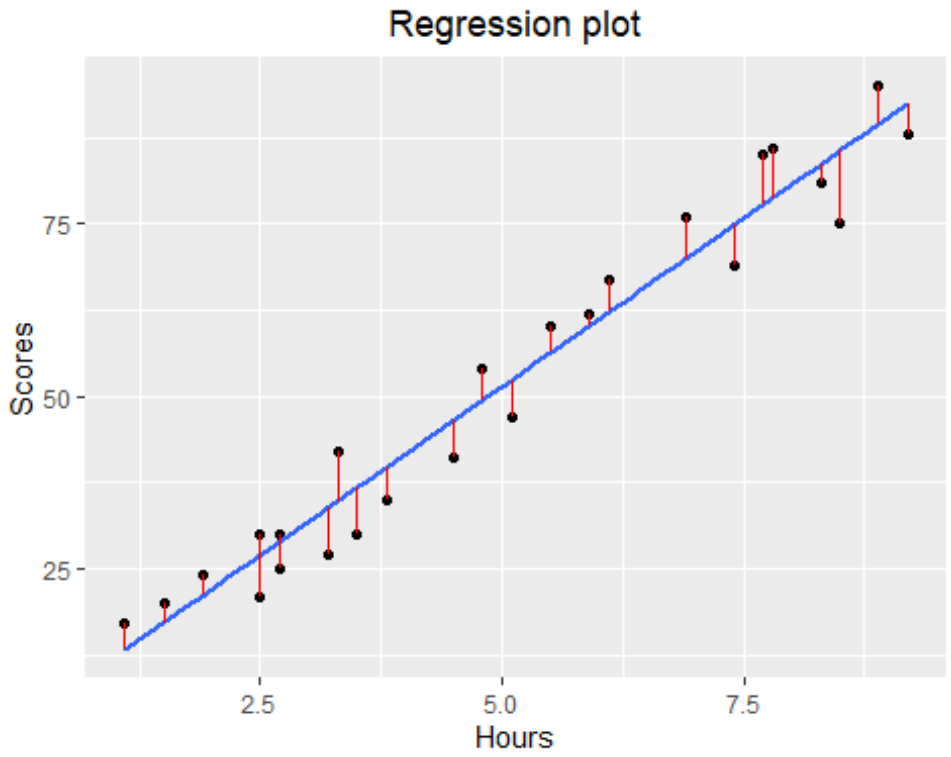
	Hours	Scores(Observed)	Expected	Residual	Standard Residual
1	2.5	21	26.9	-5.92	-1.1
2	5.1	47	52.3	-5.34	-0.973
3	3.2	27	33.8	-6.77	-1.25
4	8.5	75	85.6	-10.6	-2.01
5	3.5	30	36.7	-6.7	-1.23
6	1.5	20	17.1	2.85	0.543
7	9.2	88	92.4	-4.42	-0.858
8	5.5	60	56.3	3.75	0.684
9	8.3	81	83.6	-2.62	-0.496
10	2.7	25	28.9	-3.88	-0.72

We observe that the fitted values and the observed values differ on an average of 1.83 marks. We observe that the standard-residuals are very small which indicate that our model is very useful in prediction.

```
library(tidyverse)
```

```
ggplot(model_diagnostics, aes(dat$Hours, dat$Scores)) +
  geom_point() +
  stat_smooth(method = lm, se = FALSE) +
  geom_segment(aes(xend = dat$Hours, yend = .fitted), color = "red", size =
                 0.3)+xlab("Hours")+ylab("Scores")+ggtitle("Regression
plot")+theme(plot.title = element_text(hjust = 0.5))
```

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

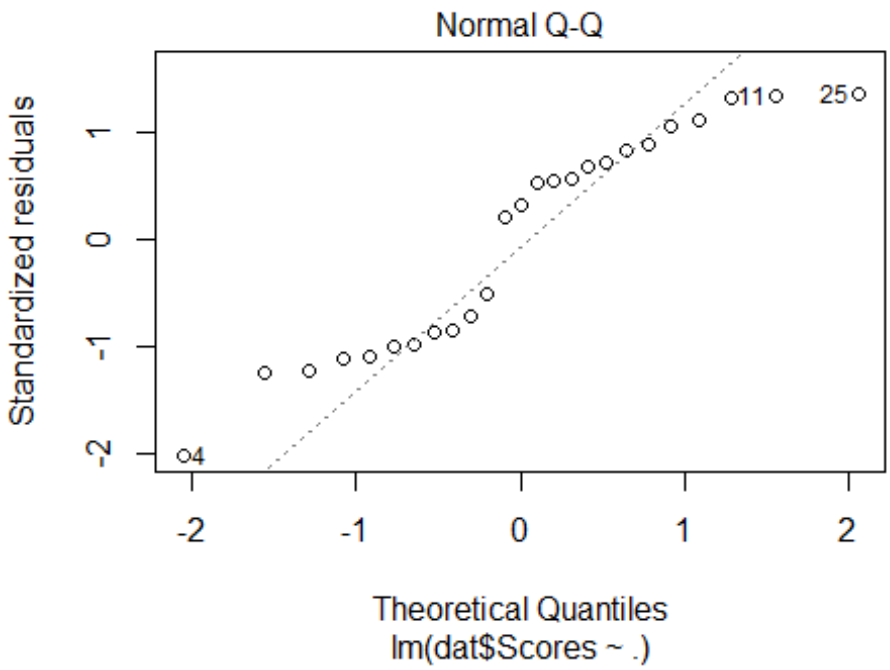


The plot shows the residuals or the difference between observed and expected values. The black dots are the observed values of the scores obtained by students and the blue line is the fitted regression line and the red verticle lines are showing the residuals or the deviation between the observed and expected values of response variable.

3**) Testing the regression assumption of normality**

In simple regression, we assume that the error or the residual is nromally distributed. We test the normality by plotting a Normal Q-Q (Quantile-Quantile) plot graphical technique for determining if two data sets come from populations with a normal distribution

```
plot(mod)
```



We observe that the plot is not along the dotted line or the standardized plots dont form a staright line. This means that the graph is not normally distributed rather is slightly negatively skewed as most data points are in the upper half. Observations 11,25 and 4 are outliers as their standard residuals are unusally large.

As the errors are not normally distributed, the response variable should be transformed such a way that the errors should be approximately normally distributed.

Conclusion:

1. We obtain the model **Y = 2.4837+ 9.7758*X** where Y is the number of marks obtained and X is the number of hours studied by student.

2. We observe that resiudals are not normally distributed and hence the response variable should be transformed such a way that the errors should be approximately normally distributed. When errors are not normally distributed, estimations are not normally distributed and we can no longer use p-values to decide if the coefficient is different from zero. In short, if the normality assumption of the errors is not met, we cannot draw a valid conclusion based on statistical inference in linear regression analysis. Hence, after transformation, a new model is requried to be constructed.