“Die Another Day (DAD)” Data Modeling

Lecture/Practical 23

29/09/2021

Case description

- Dr. X, CEO of a famous multispecialty hospital “Die Another Day (DAD)” in Tamil Nadu discussed some problems in health care sector with one of the analyst in a leading TBS consultancy services. Dr. X mentioned that the problem he is facing currently is package pricing problem. This Analyst is aware of the popular Business Model in many hospitals to quote package prices (flat rate) for treatments. i.e., Irrespective of expenses and the duration of the treatment, the patient would pay only the agreed price since it was a contract between the patient and the hospitals. Use data on “DAD” file which gives the body weight of patients and their treatment cost, can you help Dr. X to take a decision on package pricing by answering the following questions.

Try to attempt the following while building the model

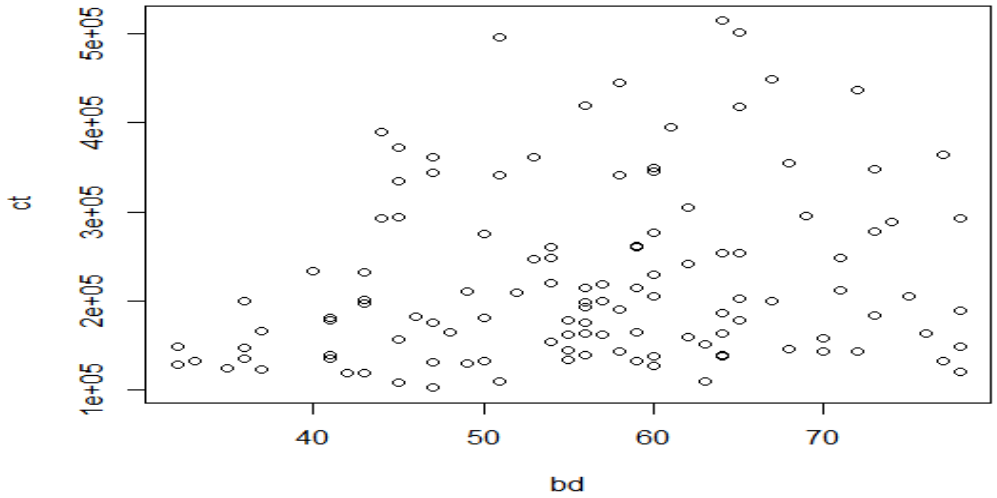
- **a)** Is there a statistical evidence to support that the cost of treatment (response) and body weight(predictor) are related? Support your answer by drawing a scatter plot?
- Build a Simple Linear Regression (SLR) model which gives the relationship between the cost of treatment (response) and body weight (predictor).
- **b)** Could you see any influential observations whose value is more than 2.5 while building the model? Remove maximum 2 such observations (one at a time) and also report the same.
- **c)** Obtain the Normal probability plot and residual versus fitted plot to check the assumptions of building a regression model? Does the assumptions of linearity, independence, homoscedasticity and normality of errors met? Comment.

Try to attempt the following while building the model (contd…)

- **(d)** Try a transformation i.e., natural logarithm only on the response variable and run a SLR with the transformed Y versus X. Obtain the plots in (c). Does the assumptions improved?
- **(e)** Write down the estimated regression equation obtained in (d)
- **(f)** What will be the average difference in cost of treatment for patient whose weight is 50 and patient whose weight is 51?

a) Is there a statistical evidence to support that the cost of treatment (response) and body weight(predictor) are related? Support your answer by drawing a scatter plot?

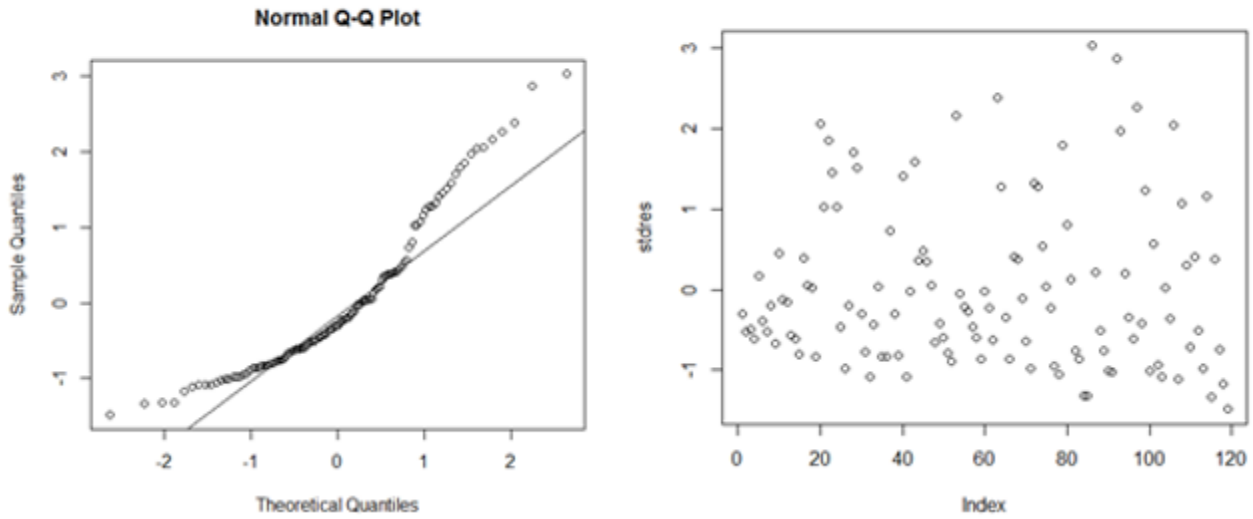
- (a) There is a moderate positive relation between cost of treatment(ct) and body weight (bd) which can be seen in the scatter plot



Build a Simple Linear Regression (SLR) model which gives the relationship between the cost of treatment (response) and body weight (predictor).
b) Could you see any influential observations whose value is more than 2.5 while building the model? Remove maximum 2 such observations (one at a time) and also report the same.

- (b) Yes. There are influential observations present .The following observations are removed (one at a time).
- # first run 2.94R. Deleted and run the regression
- #second run 3.026R. Deleted and run regression.

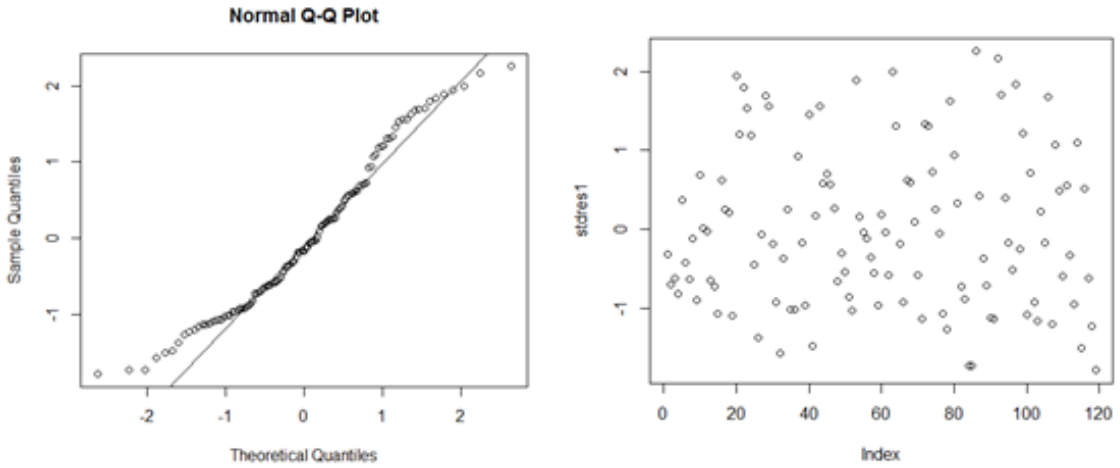
c) Obtain the Normal probability plot and residual versus fitted plot to check the assumptions of building a regression model? Does the assumptions of linearity, independence, homoscedasticity and normality of errors met? Comment.



- The assumption of normality and homoscedasticity of residual variance is not meeting. Cannot comment on linearity and independence assumptions now.

(d) Try a transformation i.e., natural logarithm only on the response variable and run a SLR with the transformed Y versus X. Obtain the plots in (c). Does the assumptions improved?

(d)



The normality plot and residual versus fitted plot seems to be improving after transformation. Now the assumptions can be considered met for building a regression model.

(e) Write down the estimated regression equation obtained in (d).

```
(e) Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.778948   0.177335  66.422   <2e-16 ***
bd           0.007785   0.003086   2.522    0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3898 on 117 degrees of freedom
Multiple R-squared:  0.05157,  Adjusted R-squared:  0.04346
F-statistic: 6.361 on 1 and 117 DF,  p-value: 0.01301
```

- The estimated model developed in (d) is $\ln(\widehat{ct})$=11.778948+0.007785(bd).

OR

$$\widehat{ct} = e^{11.778948+0.007785(bd)}.$$

(f) What will be the average difference in cost of treatment for patient aged 50 and patient aged51?

- From (e), $\widehat{ct} = e^{11.778948+0.007785(bd)}$.
- That is $\widehat{ct1} = e^{11.778948+0.007785(50)}$

  = 192566.73
- Similarly is $\widehat{ct2} = e^{11.778948+0.007785(51)}$

  = 194071.71

Hence average difference in cost of treatment for patient weight 50 and patient weight 51 is (194071.71-192566.73) =1504.98(approximately).