Linear relationship between anxiety and satisfaction levels in patients

Pranav Gopalkrishna, 1940223

2021-08-04

setwd("/Users/pranav/Documents/Study/computerScience/programming/r/rPrograms/linearRegression")

# DATA

Not much is mentioned about this data set. But from what can be learned and assumed from the given data, the following data presents observations made on a sample of patients. The fields of observation include

* Age of various patients
* Severity (of their illness, presumably)
* Their measured / self-indicated anxiety levels (in general, presumably)
* Their measured / self-indicated satisfaction levels (in general, presumably)

In this assignment, I am going to study the linear relationship, if it exists, between the two variables, ***anxiety levels*** and ***satisfaction levels*** of the patients.

completeData = read.csv("patientSatisfaction1.csv")  
head(completeData)

## Satisfaction Age Severity Anxiety  
## 1 68 55 50 2.1  
## 2 77 46 24 2.8  
## 3 96 30 46 3.3  
## 4 80 35 48 4.5  
## 5 43 59 58 2.0  
## 6 44 61 60 5.1

# IDENTIFYING INDEPENDENT AND DEPENDENT VARIABLES

Since satisfaction is generally considered desirable in sentient beings, especially in ailing patients, I assume satisfaction to be the final variable of study, meaning that the aim of any study involving these patients and variables would be to see the effect of one or more factors on their satisfaction levels. On this basis, I conclude that the satisfaction level is the dependent variable or response, while the anxiety level is the independent variable or factor. Hence, we denote anxiety level with *x* and satisfaction level with *y*.

x = completeData$Anxiety  
y = completeData$Satisfaction

# SCATTER PLOT & CORRELATION

Here, we plot satisfaction levels with respect to anxiety levels, with the former denoted using the y-axis and the latter denoted using the x-axis. We aim to observe any visual indication of an association between the two variables.

plot(x, y,  
 type = "p",  
 main = "Satisfaction with respect to anxiety levels",  
 xlab = "Anxiety",  
 ylab = "Satisfaction",  
 col = "blue",  
 pch = 16,  
 las = 1)



# type => the kind of plot. "p" means points, "l" means lines  
# col => colour  
# pch => point character type  
# las => the type of orientation of the labels on the axes

In the above scatter plot, we roughly see a negative correlation between the two variables, and while it does not seem very strong, it is consistent on average. If anxiety in fact acts as a factor for satisfaction, which is at least sometimes the case in real life, this indicates that lower anxiety levels tend to result in more satisfaction.

## CORRELATION

We will now confirm the level of association between anxiety and satisfaction levels.

myData = data.frame(x, y)  
cor(myData)

## x y  
## x 1.0000000 -0.5127287  
## y -0.5127287 1.0000000

Here, we see a negative correlation between anxiety and satisfaction levels, as previously indicated by the scatter plot. However, the correlation coefficient of -0.5127287 indicated that there is a moderate association, neither strong nor weak, indicating one or more of the following...

* Anxiety level may not be a strong factor for satisfaction level
* Anxiety level may not be the most important factor for satisfaction level
* There may be other significant contributing factors

# ESTIMATED LINEAR REGRESSION MODEL

Now, taking anxiety level to be the independent variable, we use the "lm" function in the "stats" module in R to estimate the regression coefficients (intercept and slope).

library(stats)

myLinRegModel = lm(y ~ x)  
myLinRegModel

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Coefficients:  
## (Intercept) x   
## 90.997 -6.174

y\_predicted = fitted.values(myLinRegModel)

The other more obvious way to get fitted or predicted y values would be to use the equation is...

y\_predicted = 90.997 - 6.174\*x

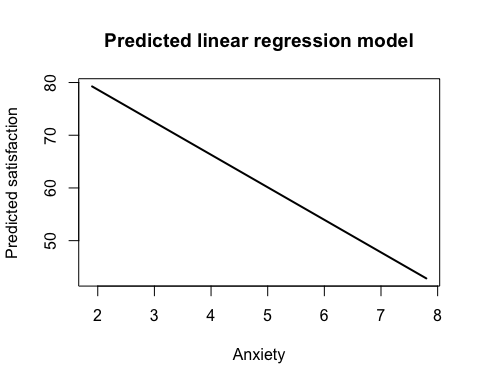
The difference is using the function, every predicted y-value is attached to an index. Furthermore, using the function demonstrably yields more accurate results, due to the lack of rounded up estimates for the coefficients.

**EXPLANATION OF CLAIMS**

You will see that the predicted y-values obtained from these two methods are very close but not the same. And, since the sum for the function-obtained values equals the sum of observed y-values, I conclude that using the function yields more accurate results. Note that these findings are a result of comparison I performed personally.

## PLOTTING THE ESTIMATED REGRESSION LINE ALONE

plot(x, y\_predicted,  
 type = "l",  
 main = "Predicted linear regression model",  
 xlab = "Anxiety",  
 ylab = "Predicted satisfaction",  
 col = "black",  
 lwd = 2)



# lwd => line width

## PLOTTING THE REGRESSION LINE WITH THE SCATTER PLOT

Now, we will graphically see how our estimated regression model fits with our sample's data i.e. how the estimated values of satisfaction levels given the anxiety levels measure compared to the observed values of satisfaction levels given the anxiety levels.

plot(x, y,  
 type = "p",  
 main = "Satisfaction with respect to anxiety levels",  
 xlab = "Anxiety",  
 ylab = "Satisfaction",  
 col = "blue",  
 pch = 16,  
 las = 1)  
abline(lm(y ~ x), lwd = 2)



# This function adds one or more regression lines through the current plot.

# COMPARING SUMS OF PREDICTED AND OBSERVED VALUES

In a correctly formed model, the sum of predicted values of *y* must equal (or be very close to) the sum of observed values of *y*.

sum(y\_predicted)

## [1] 1668

sum(y)

## [1] 1668

Hence, we have substantial confirmation that our model was correctly formed.

# OBTAINING RESIDUALS

In a correctly formed model, the sum of error terms i.e. the residuals i.e. the differences between observed and predicted values must be zero (or close to zero).

obtainedResiduals = residuals(myLinRegModel)

The other, more obvious way to derive residuals is to simply get the difference

residuals = y - y\_predicted...

sum(y - y\_predicted)

## [1] -2.842171e-14

**EXPLANATION OF CLAIMS**

You will see that the residuals obtained from these two methods are very close but not the same. And, in this case at least, since the sum for the function-obtained residuals is two times smaller than the sum of simple difference obtained residuals, I assume that using the function yields more accurate results. Note that these findings are a result of comparison I performed personally.

In any case, we see that the sum of residuals is extremely close to zero for most practical purposes involving patients and their anxiety and satisfaction levels. Hence, we have even more substantial confirmation that our model was correctly formed.