Linear relationship between anxiety and satisfaction levels in patients

Testing the obtained linear regression model's significance

Pranav Gopalkrishna, 1940223

2021-08-12

# Setting work directory

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

# Data set

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

# ESTIMATING THE LINEAR REGRESSION MODEL

To estimate the linear relationship between anxiety and satisfaction in general using the sample data, we use the lm (linear model) function.

library(stats)  
# 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

myLinRegModel = lm(y ~ x)

# TESTING THE SIGNIFICANCE OF REGRESSION PARAMETERS

Now, our aim is to estimate, given a 5% significance level[[1]](#footnote-1)

**HYPOTHESES (for beta 1)**

# H\_0: beta 1 = 0 (change in x has no linear effect on y)  
# H\_0: beta 1 =/= 0 (change in x has a linear effect on y)

**HYPOTHESES (for beta 0)**

# H\_0: beta 0 = 0 (no true mean effect)  
# H\_0: beta 0 =/= 0 (some true mean effect)  
#

## SUMMARY OF ESTIMATED LINEAR REGRESSION MODEL

summary(myLinRegModel)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -35.648 -15.274 3.291 13.613 30.143   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 90.997 9.259 9.828 1.06e-09 \*\*\*  
## x -6.174 2.156 -2.864 0.00877 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.63 on 23 degrees of freedom  
## Multiple R-squared: 0.2629, Adjusted R-squared: 0.2308   
## F-statistic: 8.203 on 1 and 23 DF, p-value: 0.008773

R-squared (i.e. ratio between the regression sum of squares i.e. the variation from the mean due to regression and the total sum of squares i.e. the variation of the actual data from the mean) measures the proportion of variation explained by our estimated regression model.

The more it explains, it means the better the model fits the data. R-squared value is 0.2629, meaning only 26.29% of the variation is explained by our estimated regression model.

Pr(>|t|) is the p-value. It denotes the probability that the calculated t-value (estimated for the population) is lesser than the table t-value. Hence, it denotes the probability that the null hypothesis is true. For intercept and x-coefficient i.e. slope, Pr(>|t|) is less than the significance level 0.05. Hence, we reject the null hypotheses for both beta 1 and beta 0, and conclude that both parameters are significant, when estimated for the population. In other words, we may conclude that

* There is some linear relationship between x and y in the population
* y exists in some amount, even when unaffected by x

These conclusions make our estimation of their linear relationship more meaningful, since we can say (assuming the population is normally distributed) with 95% confidence that a linear relationship exists.

(This means that if we take samples with similar or the same properties as the one we have taken, for around 95% of of the cases, we can expect a linear relationship between x and y in the population)

### Calculating the 95% confidence intervals for beta 1 and beta 0

confint(myLinRegModel)

## 2.5 % 97.5 %  
## (Intercept) 71.84277 110.150910  
## x -10.63363 -1.714717

Hence, 95% of the cases, the model intercept will lie between 71.84277 and 110.150910, and the slope will lie betweeen -10.63363 and -1.714717.

## PERFORMING ANOVA ON REGRESSION MODEL

ANOVA is used conclude whether the means of different classes/groups regarding a certain characteristic can be said to be equal in the population, and if not, which classes/groups can be said to differ significantly in terms of their means. ANOVA is always calculated using a sample In regression model, the different classes, or levels, are the different values of the independent variable in the sample, x i.e. anxiety level in this case. Each x-value has two replications, namely the actual y-value and the predicted y-value. If the means for all these levels i.e. x-values can be said to be equal in the population, this indicates weak or no linear relationship between x and y in the population. Hence, ANOVA on a linear regression model helps decide whether the model is significant i.e. whether the variables linearly related to some degree. This serves the same purpose as our previous hypothesis testing. Here the hypotheses are

# H\_0: Means are equal => no linear relationship => beta 1 = 0  
# H\_1: Means are different => some linear relationship => beta 1 =/= 0  
anova(myLinRegModel)

## Analysis of Variance Table  
##   
## Response: y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## x 1 2847.4 2847.38 8.203 0.008773 \*\*  
## Residuals 23 7983.7 347.12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Pr(>F) is the p-value. It denotes the probability that the calculated F-value (estimated for the population) is lesser than the table F-value. Hence, it denotes the probability that the null hypothesis is true. In our case, Pr(>F) is less that the significance level 0.05. Hence, we may reject H\_0 and conclude that the model is significant. Hence, we may conclude that the model predicts the responses to some degree. In our case, that would mean that we may conclude that the satisfaction affects anxiety to some degree.

1. ... meaning we exclude any result that falls in the range of values in the assumed population distribution (assuming the population is normal) that is the 5% least occurring values. [↑](#footnote-ref-1)