regression

Ata COLAK

2023-07-28

Regression

Introduction

This report shows a few examples of regression, specifically linear regression using R.

The report focuses on applying linear regression using R to analyze the relationship between education and GDP by fitting a linear model, to understand how changes in GDP might be associated with changes in education. Additionally, I visualized the relationship using a scatter plot and determine the coefficient of determination to assess the model's goodness of fit.

The dataset used for this report is "Sustainable Society Indices 2016 Dataset". The dataset can be accessed using the link below:

https://github.com/gagolews/teaching-data/blob/master/marek/ssi_2016_indicators.csv

First Part - Loading Dataset

In the first example below, the csv file is loaded and read by my device, and prints the head of the dataset to show that it has been read successfully.

```r
SSI <- read.csv(paste0("https://raw.githubusercontent.com/gagolews/",
"teaching-data/master/marek/ssi_2016_indicators.csv"),
comment.char="#")
head(SSI, 3)
```

```
##   Country SufficientFood SufficientWater SafeSanitation Education HealthyLife
## 1 Albania          10.00            9.51           9.32  8.844812    8.133333
## 2 Algeria          10.00            8.40           8.74  8.096907    7.716667
## 3  Angola           8.58            4.86           5.11  6.709523    4.316667
##   GenderEquality IncomeDistribution PopulationGrowth GoodGovernance
## 1           7.01           8.381965         8.230960       4.955972
## 2           6.32           5.545405         4.087104       3.278008
## 3           6.37           3.292725         1.000000       2.976937
##   Biodiversity RenewableWaterResources Consumption EnergyUse EnergySavings
## 1     5.509341                 9.56590    5.540515     8.386      2.325103
## 2     6.595260                 3.30800    6.736071     7.346      1.000000
## 3     4.101328                 9.95245    7.566049     8.788      3.560209
##   GreenhouseGases RenewableEnergy OrganicFarming GenuineSavings      GDP
## 1          8.5757        2.728559       1.095706       4.034553 5.462496
## 2          6.8426        1.000000       1.003808       9.414452 6.292930
## 3          9.2035        5.172046       1.009591       3.571235 3.910628
##   Employment PublicDebt
## 1   1.998876   2.615133
## 2   3.867410   9.577604
## 3   5.066170   2.669161
```

Second Part - Fitting a Linear Model for Education - GDP and Explanation

In order to find the fitted model equation, we can observe the values by looking through the "coefficients" section of the list "model", where:

The (Intercept) value in coefficients indicate the estimated coefficient of the intercept, and

The GDP value in coefficients indicate the estimated coefficient of GDP.

By putting these values in y = mx = b, we receive "Education = 0.41 * GDP + 5.42", which is the estimated fitted model equation, where 5.42 indicates the estimated coefficient of intercept, and 0.41 indicates the estimated coefficient of GDP. Therefore, the full fitted model equation is:

```r
model <- lm(Education ~ GDP, data = SSI)

fitted_print <- paste("Education =", coef(model)[2], "* GDP", "+", coef(model)[1])
print(fitted_print)
```

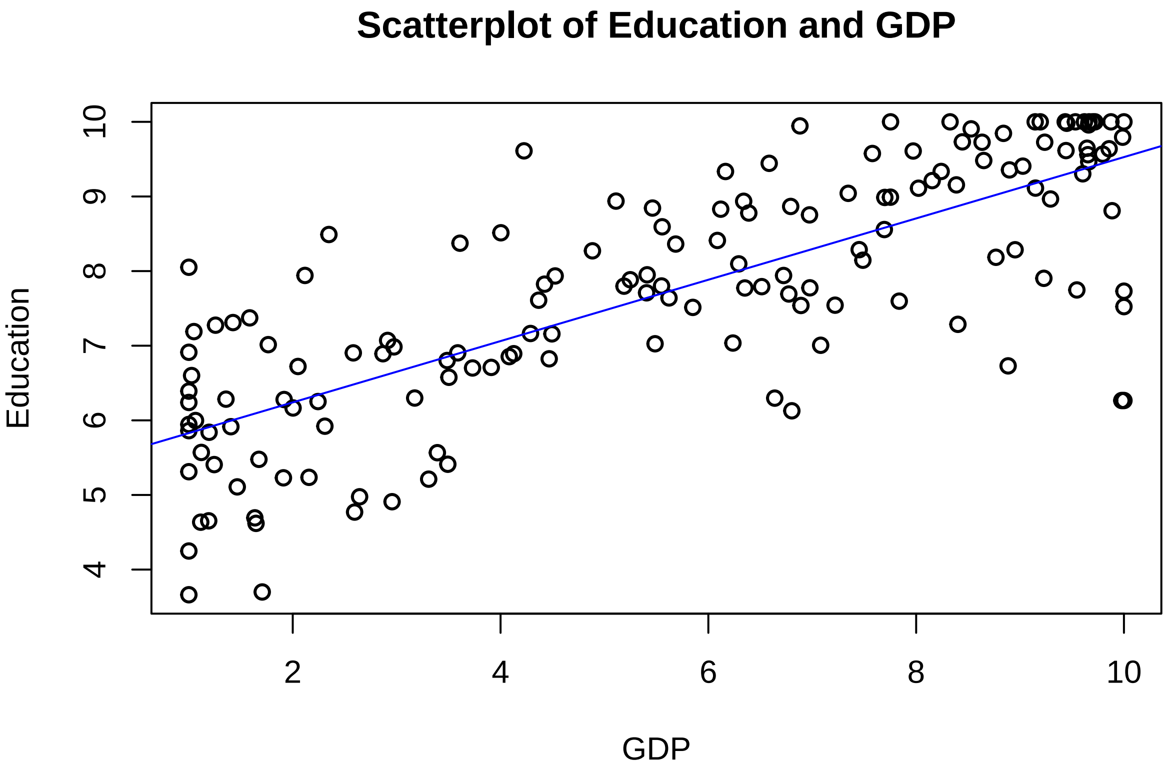
```
## [1] "Education = 0.410897386217405 * GDP + 5.41864568766329"
```

Third Part - Scatterplot With Fitted Linear Regression Line

In the third part, I have plotted GDP and education, and I have added the fitted linear regression line which is "model" as a straight line to the aforementioned scatterplot.

This scatterplot shows the correlation between GDP and education, and aims to accurately show the linear regression between the two data.

```r
plot(SSI$GDP, SSI$Education,xlab = "GDP", ylab = "Education", main = "Scatterplot of Education and GDP"
abline(model, col = "blue")
```



Fourth Part - Coefficient of Determination of Model And Interpretation

To obtain the coefficient of determination, we use the "summary()" function with "model" and ask for the "r.squared" value. Using this method, the coefficient of determination of the linear model on this dataset is:

```
summary(model)$r.squared
```

```
## [1] 0.6043174
```

According to this value, we can see that the coefficient of determination is approximately 0.6, which shows that this linear model has a moderate correlation with the dataset. A value over 0.7 is accepted as a strong correlation, therefore we can observe that this model somewhat correlates with the dataset.

Fifth Part - Prediction of Education Scores for the Given GDP Values

To predict the education scores for GDP of 2.5, 5 and 7.5, I first created a new data frame with the GDP values. Afterwards, I used the predict function to predict the education scores. And lastly, I displayed the predicted scores of education. I have created a second dataset to display the prediction scores so the result is more presentable.

```
GDP_dataframe <- data.frame(GDP = c(2.5, 5, 7.5))
predictedScores <- predict(model, GDP_dataframe)
predicted_scores_print <- data.frame(GDP_dataframe, predictedScores)
predicted_scores_print
```

```
##   GDP predictedScores
## 1 2.5        6.445889
## 2 5.0        7.473133
## 3 7.5        8.500376
```

Conclusion

In conclusion, this report focused on applying linear regression using R to analyze the relationship between education and GDP. A scatter plot was created to visualize the relationship, and the coefficient of determination was approximately 0.6, indicating a moderate correlation between education and GDP in the dataset.

While doing this report, I have learned how to create a fitted linear regression line, add a fitted linear regression line on a model, computing the coefficient of determination, and using the predict function.

What was interesting to me about this report was how simple it was to predict the approximate information without having the related information in the dataset using linear models, and a thought provoking question was how inaccurate could the predictions be if the values asked in a dataset were with very high numbers, and what a plot of the correlation between the size of values - coefficient of determination would look like.