Building a Model for weight and systolic bp of male people aging 25-30

Nithya S

2023-11-03

Introduction

A linear regression model is a statistical method that describes the relationship between a dependent variable and one or more independent variables. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables.

Objective

• Obtaining the scatter plot and interpreting.

• Finding the regression line connecting the variables systolic bp and weight. To interpret the plot, intercept term and the regression coefficients. To infer from the sign of regression coefficient.

• Obtaining the fitted values. To check if the sum of fitted values is equal to the sum of observed values.

URL for the dataset:

https://docs.google.com/spreadsheets/d/1WXzKqlu21X3qhfAWn1-FaoY3OkLI78VD/edit?usp=sharing&ouid=110138716074493614410&rtpof=true&sd=true

Data description

The weight and systolic blood pressure of 26 randomly selected males in the age group 25–30 are taken into consideration for the study. Here we will be intended to build a model and understand the relationship between the variables. We have to note that the data collected here is of male people aging from 25 to 30. So we take weight as the dependent variable and systolic bp as the independent variable.

Procedure and analysis

We import the dataset from excel in order to proceed further. We import the dataset as follows

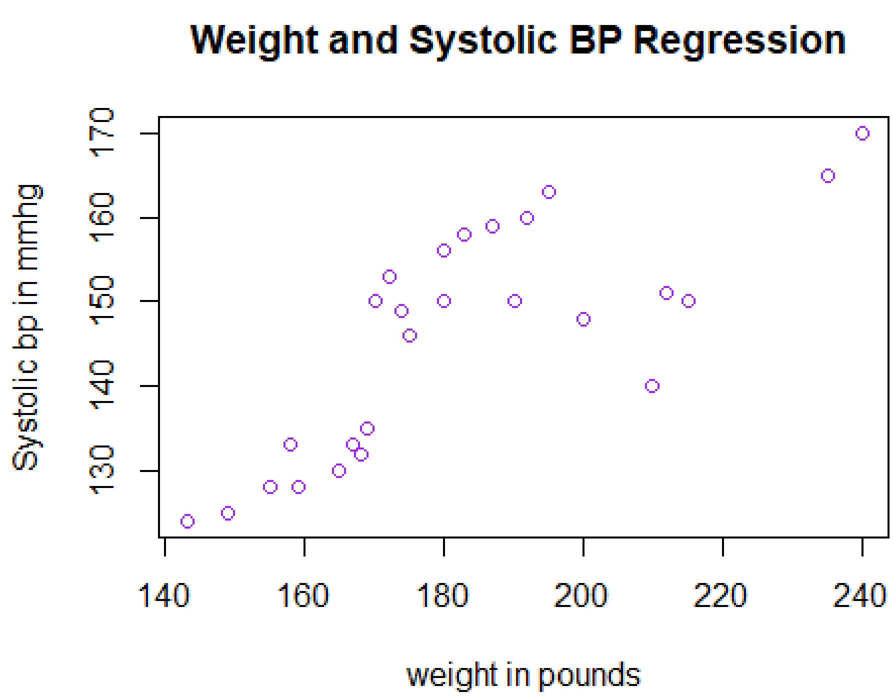
```
library(readxl)
Dataset <- read_excel("G:/My Drive/Linear Regression/Datasets/Dataset.xlsx")
View(Dataset)
attach(Dataset)
```

Now that we have attached the dataset, we now proceed to build the model.

To check the relationship between the variables under study

1. By plotting

```
plot(Dataset$Weight,Dataset$`Symbolic BP`,col="purple",main="Weight and Systo
lic BP Regression ",xlab="weight in pounds",ylab="Systolic bp in mmhg")
```



From the scatter plot we interpret as: the independent variable (X) is weight in pounds and the dependent variable (Y) is systolic bp in mmhg. The plot shows that there is a linear positive relationship between the variables under study. The relationship can be understood better numerically by calculating the Karl Pearson's co-efficient.

2.By Karl Pearson's correlation co-efficient

```
cor(Dataset$Weight,Dataset$`Symbolic BP`)

## [1] 0.7734903
```

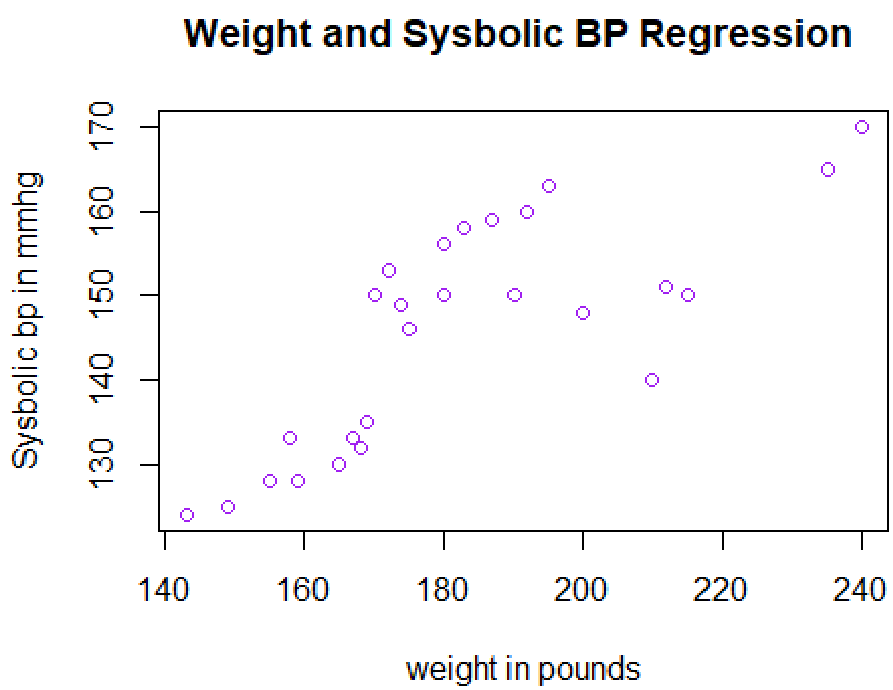
The correlation co-efficient between the variables is 0.7734903. So we say that there is a moderately positive linear relationship between the variables under study.

We give a regression equation

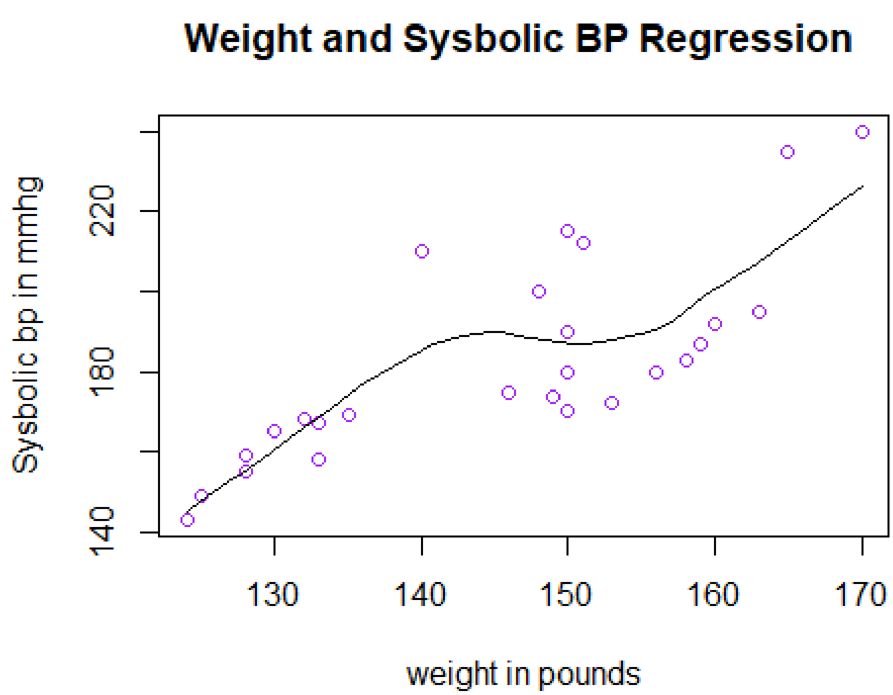
```
model=lm(Dataset$`Symbolic BP`~Dataset$Weight)
model

##
## Call:
## lm(formula = Dataset$`Symbolic BP` ~ Dataset$Weight)
##
## Coefficients:
##    (Intercept)  Dataset$Weight
##        69.1044          0.4194

summary(model)

##
## Call:
## lm(formula = Dataset$`Symbolic BP` ~ Dataset$Weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.182  -6.485  -2.519   8.926  12.143
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69.10437   12.91013   5.353 1.71e-05 ***
## Dataset$Weight  0.41942    0.07015   5.979 3.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.681 on 24 degrees of freedom
## Multiple R-squared:  0.5983, Adjusted R-squared:  0.5815
## F-statistic: 35.74 on 1 and 24 DF,  p-value: 3.591e-06
```

To plot the observations with a smooth line

```
plot(Dataset$Weight,Dataset$`Symbolic BP`,col="purple",main="Weight and Systo
lic BP Regression ",xlab="weight in pounds",ylab="Systolic bp in mmhg")
```




```
scatter.smooth(Dataset$`Symbolic BP`,Dataset$Weight,col="purple",main="Weight and Systolic BP Regression ",xlab="weight in pounds",ylab="Systolic bp in mmhg")
```



The smooth scatter plot shows the relationship of the variables again but in a informative manner. The line passing through the points gives a better understanding and we see that it is not a straight line but a tilted one.

The model here is

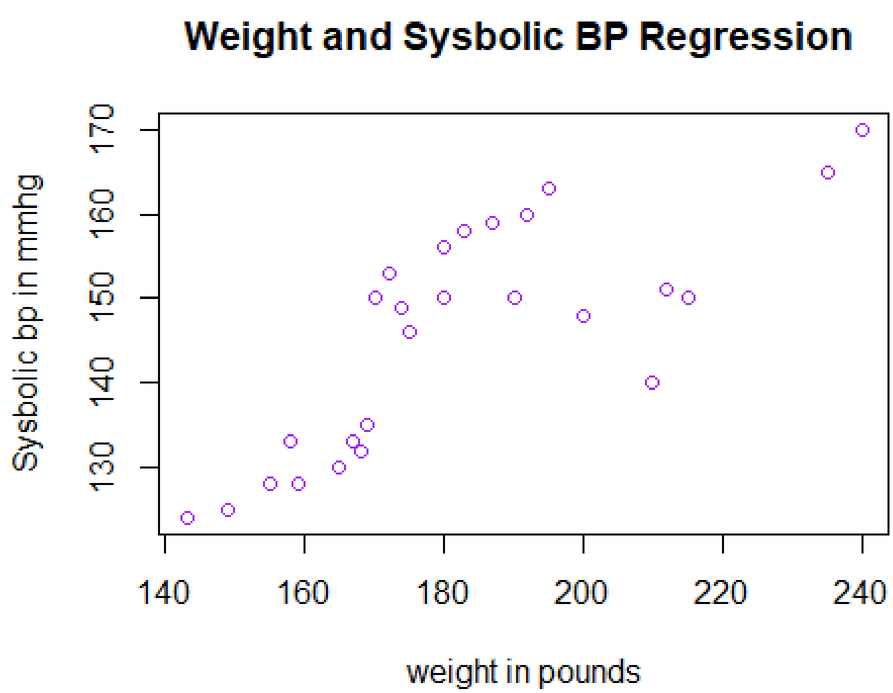
$$Y = B_0 + B_1X + U$$

From the model table, we see that the intercept parameter(B_0) is 69.1044 and the slope parameter(B_1) is 0.41942. The intercept term (B_0) depicts that if X sometimes becomes zero, the intercept is simply the expected value of Y at that value. The slope term (B_1) depicts that for every 1 unit increase in X, the value of Y increases by 0.41942

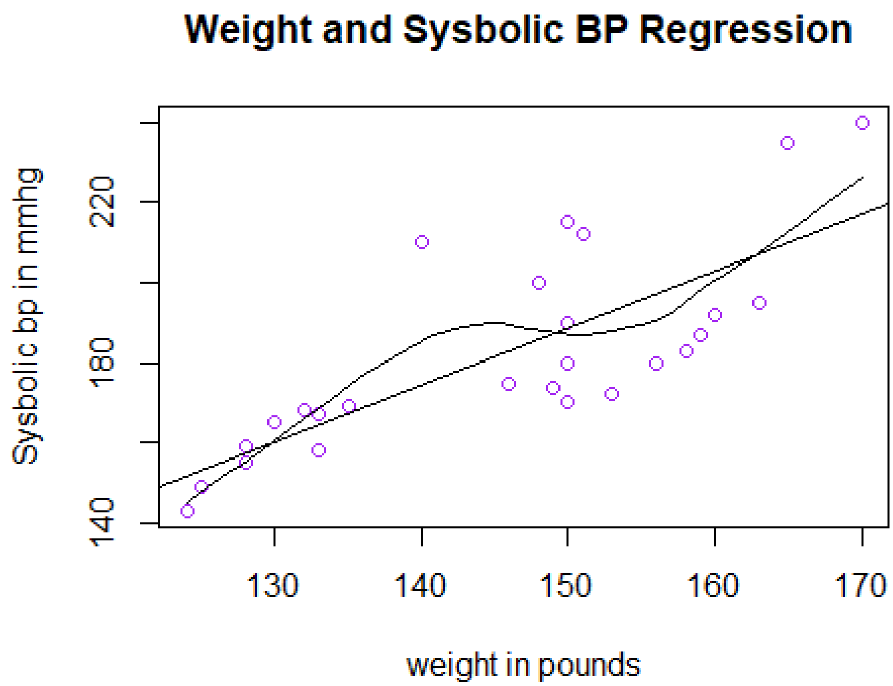
The regression co-efficient's sign can be interpreted as, since B_1 is positive the regression co-efficient is also positive because the covariance has to be positive in order for the B_1 to be positive.

Fitted lines in a scatter diagram

```
plot(Dataset$Weight,Dataset$`Symbolic BP`,col="purple",main="Weight and Systo
lic BP Regression ",xlab="weight in pounds",ylab="Systolic bp in mmhg")
```




```
scatter.smooth(Dataset$`Symbolic BP`,Dataset$Weight,col="purple",main="Weight
and Systolic BP Regression ",xlab="weight in pounds",ylab="Systolic bp in mmh
g")

abline(lm(Dataset$Weight~Dataset$`Symbolic BP`))
```



The interpretation that can be drawn from the above graph is that, the above graph is the best possible fit for the dataset. We also see that there are ,many outliers from the above graph.

To obtain the fitted values

```
Fit=fitted.values(model)
Fit

##        1        2        3        4        5        6        7        8
## 138.3079 139.1467 144.5991 134.1137 158.0204 142.5020 148.7933 157.1816
##        9       10       11       12       13       14       15       16
## 152.9874 131.5972 135.3720 139.9855 140.4050 141.2438 135.7914 139.5661
##       17       18       19       20       21       22       23       24
## 142.0826 145.8574 159.2786 150.8903 144.5991 129.0807 169.7640 167.6669
##       25       26
## 149.6321 147.5350
```

The above are the fitted values for the “Systolic bp (Y)”.

To see if the sum of the fitted values is equal to the sum of the observed values. First we sum the fitted values. Then we sum the Systolic bp values and check if they both are the same as follows.

```
sum_fit=sum(Fit)
sum_fit

## [1] 3786
```

```
sum_sysbp=sum(Dataset$`Symbolic BP`)
sum_sysbp

## [1] 3786
```

We conclude that the sum of the fitted values is equal to the sum of the observed values.

Conclusion

The above dataset is all about the weight and systolic bp of 26 males aging from 25 to 30. There is a moderately linear relationship between the variables with the correlation co-efficient as 0.7734903 The model considered here is

$$Y = B_0 + B_1X + U$$

B_0=69.1044

B_1=0.41942

U is the error term

The reason for consideration of weight as the independent variable and systolic bp as dependent variable is as follows:

1. The given dataset is of the age group 25-30, so mostly the male people will not have any health issues.
2. Although a decrease in salt intake during dieting may contribute to the blood pressure lowering effect of weight reduction. But in general bp depends on the weight of the person.

Also, the values of the evaluated co-efficients are significant and the null hypothesis is rejected.