Linear Regression

Hamid Abdulsalam

Introduction

In regression, we are interested in knowing the type of relationship that exists between variables. We can use this relationship to build a model that can be used for describing, controlling, and predicting.

In a Simple regression analysis, if we know the value of the independent variable, there is a model that will give us an estimate for the value of the dependent variable.

The stronger the relationship between the independent and dependent variables, the better estimates the model will give.

Introduction

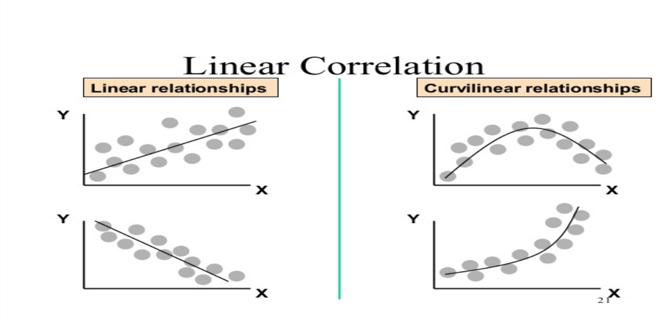
A scatter plot is often used to show the relationship between two variables graphically.

Correlation analysis is used to measure strength of the linear relationship association between two variables. Correlation does not mean causation. Hence, no causal effect is implied with correlation

The type of model built depends upon the type of variables and the type of relationship that exists between the variables.

Types of Relationship

Consider the following scatterplots:



- X is called the “independent” or “predictor” variable
- Y is called the “response” or “dependent” variable

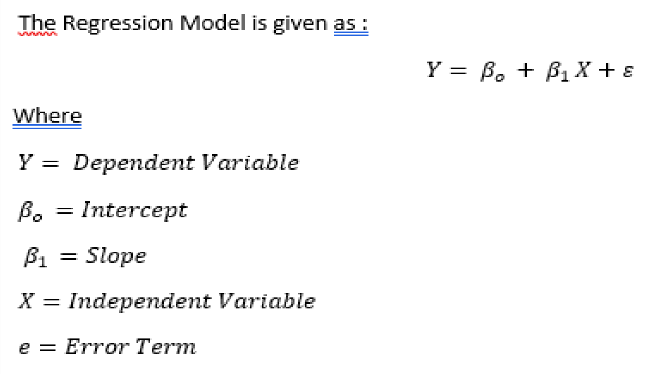
Some ssumptions of Linear Regression

- Linear relationship between dependent and independent variables (This assumption can be tested through scatter plot)
- Normality of residuals (This assumption can be checked by looking at a histogram or a Q-Q-Plot of the residuals. Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test).
- The data come from a random sample of size n from the population of interest or a randomized experiment
- Constant variance assumption

Simple Linear Regression

Simple Linear Regression

In Simple Linear Regression, there is one response variable (Y), one predictor variable (X) and the relationship between the response and predictor variable is linear.



The Regression Model is given as :

$$Y = \beta_o + \beta_1 X + \varepsilon$$

Where

$Y = Dependent\ Variable$

$\beta_o = Intercept$

$\beta_1 = Slope$

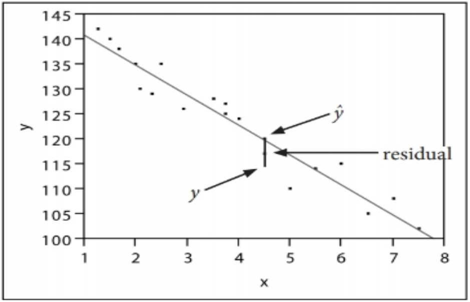
$X = Independent\ Variable$

$e = Error\ Term$

Residual

The residual also known as the error term is the difference between the observed value and the predicted value. In regression modelling, we are always interested in the model that minimises the residual sum of squares (RSS)

Residual = observed value − predicted value = $y - \hat{y}$



Case Study 1

As a data scientist, a client approached you to analyze his business data. He wants to know if there is a significant relationship between the Age and the Income of his customers who purchased his treadmill products.

We start by loading the data set into R.

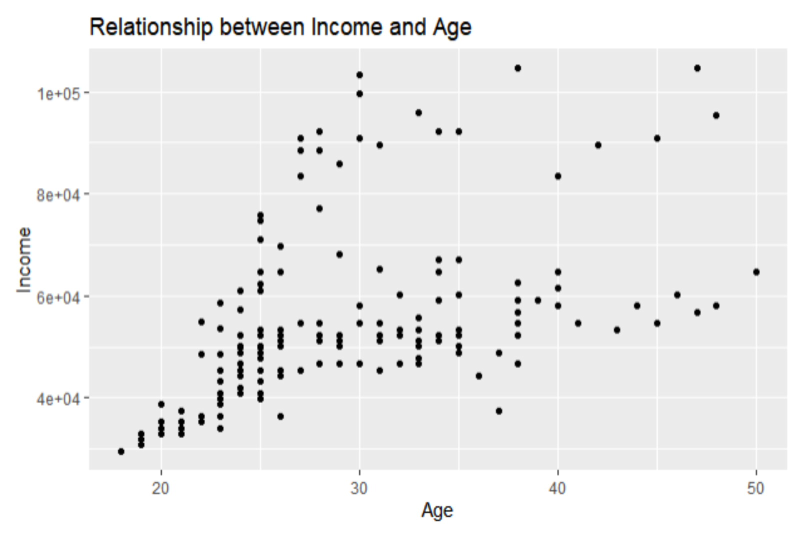
```r
dt<-read.csv("data/Gym.csv", header = T)
```

Case Study 1

Let’s examine the relationship between Age and Income

```r
library(ggplot2)
p1<-ggplot(dt, aes(x=Age, y=Income))+geom_point()+
  labs(title = "Relationship between Income and Age",
       xlab="Age", ylab="Income")
```

## Case Study 1


Relationship between Income and Age
Income
1e+05
8e+04
6e+04
4e+04
Age
20
30
40
50

Case Study 1

What do you observe in the scatter plot? You can see that there is a moderate positive linear relationship between the Age and the Income of the Customers. As their age increases so does their income also increases.

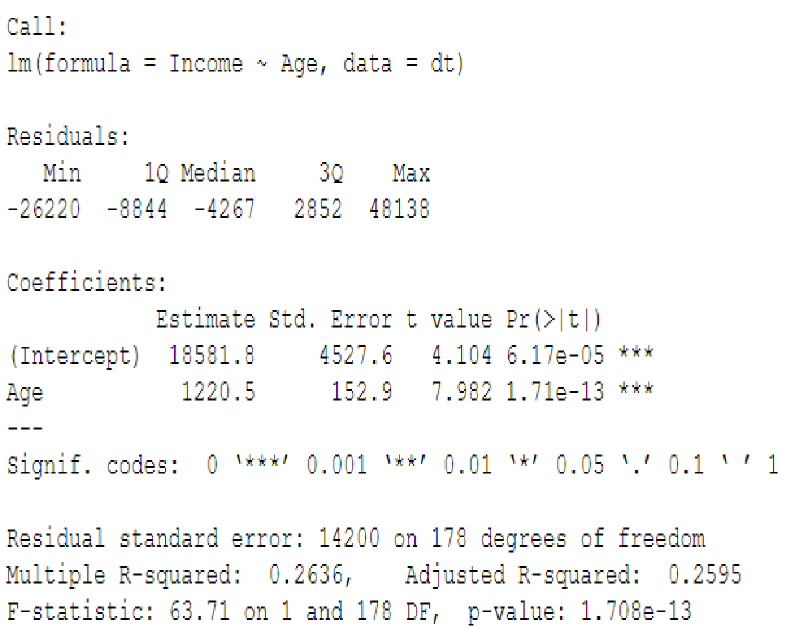
Since we have established that there is a linear relationship between Age and Income. Let's see if there is a causal relationship using regression modeling, does their age affects to their Income?

Case Study 1

To fit a linear regression model in R, we use the lm function

```r
m1<-lm(Income~Age, data = dt)
```

Case Study 1



```
Call:
lm(formula = Income ~ Age, data = dt)

Residuals:
   Min     1Q Median     3Q    Max 
-26220  -8844  -4267   2852  48138 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  18581.8     4527.6   4.104 6.17e-05 ***
Age           1220.5      152.9   7.982 1.71e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14200 on 178 degrees of freedom
Multiple R-squared:  0.2636,	Adjusted R-squared:  0.2595 
F-statistic: 63.71 on 1 and 178 DF,  p-value: 1.708e-13
```

Case Study 1

The output depicts that the linear equation is

$y = 18581.8 + 1220.5(\text{Age})$

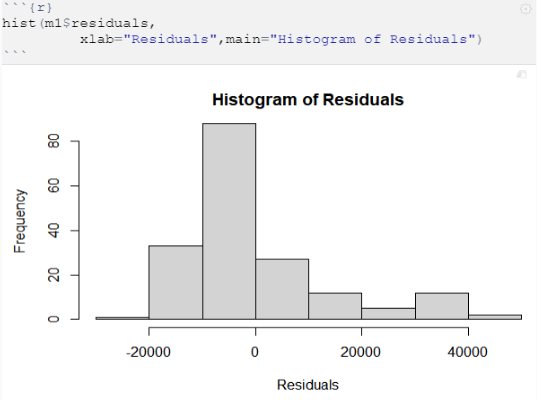
Since the P-value of 1.71e-13 of the coefficient of Age (slope) is less than 0.05, this indicates that the Age has significant effect on Income.

The intercept is also significant at 5% level of significance.

To interpret the coefficient of Age (slope), for a unit increase in the Age, the Income increases by 1220.5.

The Adjusted R-squared is interpreted as the amount of variation of the dependent variable that can be explained by the independent variable. For the model, the 25.95% of the total variation in Income can be explained by the Age.

Normality of Residuals of our model



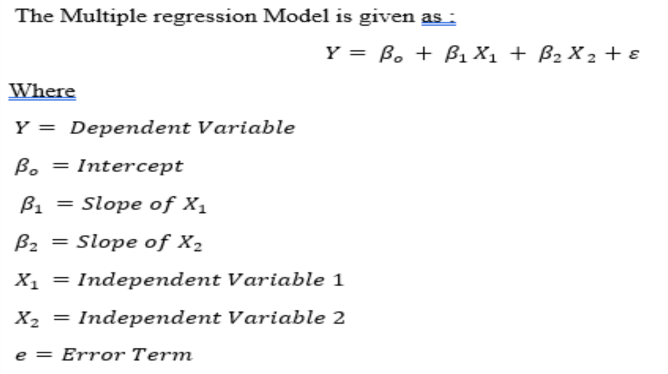
~~~
```{r}
hist(m1$residuals,
          xlab="Residuals",main="Histogram of Residuals")
```
~~~

The residuals are skewed to the right, which indicates that the assumption of normality of residuals has been violated. We can rectify this by taking log transformation of our variables.

Multiple Linear Regression

Multiple Linear Regression

A simple linear regression is for a single response variable, y, and a single independent variable, x. Multiple linear regression is used when you have more than one independent variable.



The Multiple regression Model is given as :

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Where

$Y = Dependent\ Variable$

$\beta_o = Intercept$

$\beta_1 = Slope\ of\ X_1$

$\beta_2 = Slope\ of\ X_2$

$X_1 = Independent\ Variable\ 1$

$X_2 = Independent\ Variable\ 2$

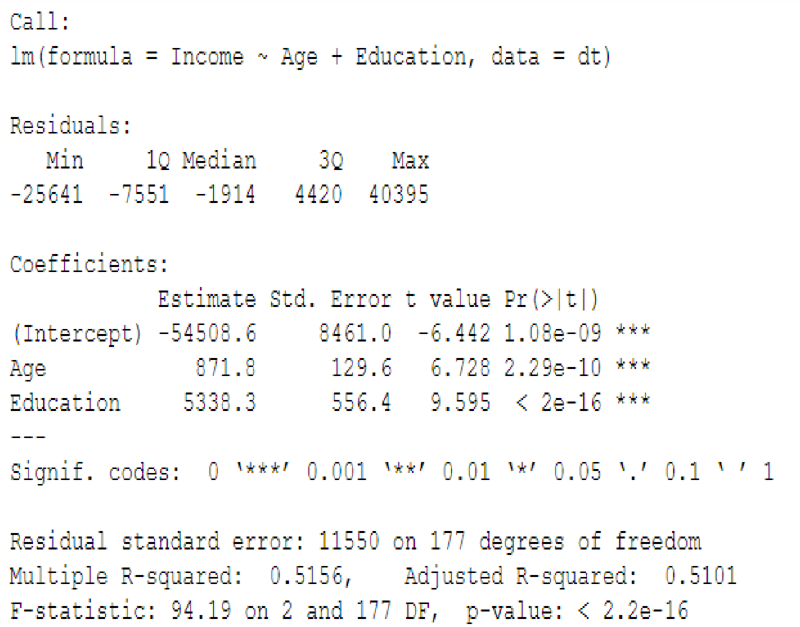
$e = Error\ Term$

Multiple Linear Regression

As we can see from the simple regression model in case study 1, we have a weak model with just 25.95% of the total variation in Income explained by the Age. We can improve this model through multiple regression model by adding more variable in the model. Let's add Education level to the model, we want to know if their Age and education level have affect their income.

```
m2<-lm(Income~Age+Education, data = dt)
```

Case Study 2



```
Call:
lm(formula = Income ~ Age + Education, data = dt)

Residuals:
   Min     1Q Median     3Q    Max
-25641  -7551  -1914   4420  40395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -54508.6     8461.0  -6.442 1.08e-09 ***
Age            871.8      129.6   6.728 2.29e-10 ***
Education     5338.3      556.4   9.595  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11550 on 177 degrees of freedom
Multiple R-squared:  0.5156,    Adjusted R-squared:  0.5101
F-statistic: 94.19 on 2 and 177 DF,  p-value: < 2.2e-16
```

Case Study 2

Look at the Adjusted R-squared, by adding education to the model,our model improved by 25%. Hence, far better than our previous model.

The output depicts that the linear equation is

y= -54508.6 + 871.8(Age) + 5338.3(Education)

Since the P-value of 2.29e-10 of the coefficient of Age is less than 0.05, this indicates that the Age has significant effect on their Income.

Also the P-value of the coefficient of Education is less than 0.05, this indicates that the Education also has significant effect on their Income.

Case Study 2

The intercept is also significant at 5% level of significance.

To interpret the coefficient of Age , for a unit increase in the Age, the Income increases by 871.8.

To interpret the coefficient of Education , for a unit increase in the Education level, the Income increases by 5338.3

Hence, the higher the Age and Education, the higher their income.

For the multiple regression model, 51.01% of the total variation in Income can be explained by Age and Education. This indicates an improved model.

Application of the Model

So let's say the client is interested in using the model for prediction. He wants to predict what the expected income of a customer who is 50 years old and has spent 15 years on Education.

To do this in R, we can obtain both the predicted value and 95% prediction interval using the Predict Function.

```r
newdata<-data.frame(Age=c(50), Education=c(15))
newdata
```

```
##   Age Education
## 1  50        15
```

```r
predict(m2, newdata = newdata, interval = 'prediction')
```

```
##        fit      lwr      upr
## 1 69156.93 45609.63 92704.22
```

Application of the Model

The predicted income is 69156.93 with a Prediction interval of 45609.63 and 92704.22. The prediction interval indicates that we are 95% confident that the person whose Age is 50 and has spent 15 years on education will have an income that falls between the range 45609.63 and 92704.22.