# Linear Regression Analysis

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line. In previous learning, you have placed the line of best fit by eye. Each student would have produced different best fitting lines, therefore different results would have been given for predictions. For better accuracy, there is a more mathematical way to establish this line.

We will use computer software to calculate this line of best fit; the Least Squares Regression Line.

The general equation of the line is 𝑦=𝑎+𝑏𝑥 where 𝑏 is the slope of the line and 𝑎 is the y-intercept (where the line crosses the *y*-axis.

The equation of the Least Squares Regression line can be produced in a calculator but for the purpose of this work, we will use computer software. i.e. EXCEL.

**Example 1:** The following table shows the average temperature in July every year.

|  |  |
| --- | --- |
| Year | Average Temp |
| 1990 | 20.8 |
| 1992 | 20.9 |
| 1994 | 20.9 |
| 1996 | 21.1 |
| 1998 | 21.2 |
| 2000 | 21.3 |
| 2002 | 21.3 |
| 2004 | 21.5 |
| 2006 | 21.9 |
| 2008 | 21.8 |
| 2010 | 22.4 |

\*enter the years into excel as 0, 2 etc…

(a) Type the data into Excel and save as a csv file.

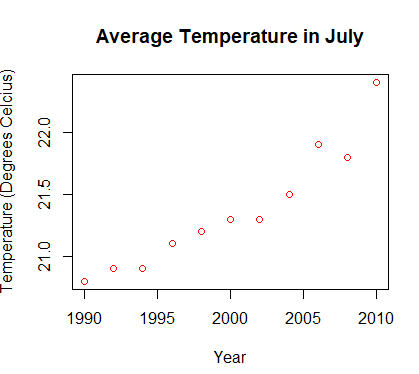
(b) Import the data into R Studio.

July<-read.csv("July.csv")

attach(July)

(c) Draw a Scatter Plot in R Studio

plot(Year,AverageTemp, main = "Average Temperature in July", ylab = "Temperature (Degrees Celcius)", col = "red")

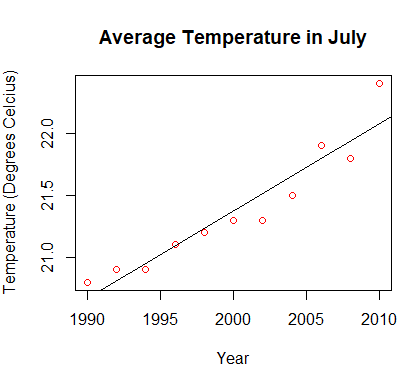


(d) Create a linear model and write the equation in the form y = mx + c where x is the year and y is the average temp. Write it below

|  |
| --- |
| lm(formula = AverageTemp ~ Year)  Coefficients:  (Intercept) Year  -119.53636 0.07045  Temperature = 0.07045(Year) – 119.53636 |

(e) Add a line of best fit to your model

abline(lm(AverageTemp~Year))



(f) Use your linear model to predict what average temp will be in 2018.

predict(lm(AverageTemp ~ Year), newdata=data.frame(Year=2018))

22.64091

DON’T FORGET TO REMOVE DATA

> detach(July)

> remove(July)

**Example 2.** The table below gives the amount of Krabby Patties made by Spongebob for each year he’s worked. Graph the data on a scatter plot, find the line of best fit, and write the equation for the line you draw.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Years worked** | 1 | 2 | 3 | 4 | 5 | 6 |
| **Patties made** | 6,500 | 7,805 | 10,835 | 11,230 | 15,870 | 16,387 |

(a) Using statistical software write the linear regression equation.

|  |
| --- |
| Krabby<-read.csv("Krabby.csv")  attach(Krabby)  lm(formula = PattiesMade ~ YearsWorked)  Coefficients:  (Intercept) YearsWorked  4035 2115  Patties = 2115(Years) + 4035 |

(b) Calculate the Correlation Coefficient (r).

|  |
| --- |
| cor.test(YearsWorked,PattiesMade)  Pearson's product-moment correlation  data: YearsWorked and PattiesMade  t = 9.0264, df = 4, p-value = 0.0008344  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  0.7934284 0.9975105  sample estimates:  cor  0.9763211 |

Coefficient of Determination

The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.

With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.

An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.

An R2 of 1 means the dependent variable can be predicted without error from the independent variable.

An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R2 of 0.20 means that 20 percent is predictable; and so on.

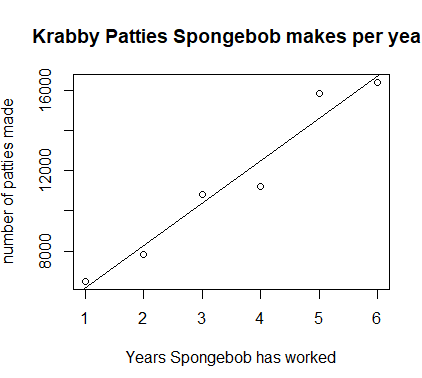
(c) Calculate the coefficient of determination.

|  |
| --- |
| R2 =0.97632112 = 0.95 |

(d) Add a line of best fit for your model

> plot(YearsWorked,PattiesMade, main = "Krabby Patties Spongebob makes per year", xlab = "Years Spongebob has worked", ylab = "number of patties made")

> abline(lm(PattiesMade~YearsWorked))



(e) Using the linear regression equation predict how many Krabby Patties he will make after working 10 years.

|  |
| --- |
| predict(lm(PattiesMade~YearsWorked), newdata=data.frame(YearsWorked=10)) |

**Exercise: Linear Regression Analysis**

1. The table below gives the amount of time students in a class studied for a test and their test scores. Graph the data on a scatter plot, find the line of best fit, and write the equation for the line you draw.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hours Studied** | 1 | 0 | 3 | 1.5 | 2.75 | 1 | 0.5 | 2 |
| **Test Score** | 78 | 75 | 90 | 89 | 97 | 85 | 81 | 80 |

(a) Using statistical software to create a linear model, compute the linear regression equation.

Hours<-c(1,0,3,1.5,2.75,1,0.5,2)

Score<-c(78,75,90,89,97,85,81,80)

lm(Score~Hours)

Call:

lm(formula = Score ~ Hours)

Coefficients:

(Intercept) Hours

76.406 5.425

Score = 5.425(Hours) + 76.406

(b) Compute the correlation coefficient.

cor.test(Score,Hours)

cor

0.7859828

(c) Compute the coefficient of determination.

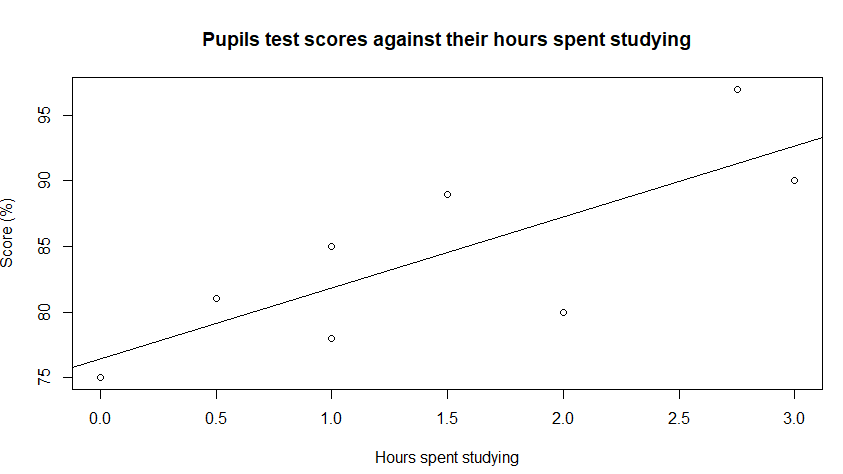
> 0.7859828^2

[1] 0.617769

(d) Create a scatter plot and add a line of best fit

plot(Hours,Score, main = "Pupils test scores against their hours spent studying", xlab = "Hours spent studying", ylab = "Score (%)")

abline(lm(Score~Hours))



(e) Using the linear regression equation predict a student’s test score if they studied for 4 hours.

> predict(lm(Score ~ Hours), newdata=data.frame(Hours=4),)

1

98.10811

2. The table below gives the estimated world population (in billions) for various years.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | 1980 | 1990 | 1997 | 2000 | 2005 | 2011 |
| **Population** | 4400 | 5100 | 5852 | 6080 | 6450 | 7000 |

(a) Using statistical software to create a linear model, compute the linear regression equation.

(b) Compute the correlation coefficient.

(c) Compute the coefficient of determination.

(d) Create a scatter plot and add a line of best fit

(e) Using the linear regression equation predict the world population in the year 2015.

3. The table below shows the income for an employee over his first 8 years of work. Use this to estimate his income for his 15th year of work.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Years** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| **Income** | 45,000 | 46,814 | 48,212 | 52,870 | 54,125 | 58,532 | 61,075 | 62,785 |

(a) Using statistical software to create a linear model, compute the linear regression equation.

(b) Compute the correlation coefficient.

(c) Compute the coefficient of determination.

(d) Using the linear regression equation predict his income for his 15th year of work.

4. For a piece of coursework, Tom investigates cars passing their school from 3pm to 4pm.

Tom records the make and model of each in his sample.

For each car, he found the figures for the fuel economy in miles per gallon (F) and the CO2 emissions in grams per kilometre (C).

He plotted F against C on this scatter diagram.

*Chart

Description automatically generated*

(a) Calculate the equation of the regression line of F on C. (Using statistical software)

(b) Draw the regression line.

(c) Calculate the correlation.

(d) You have to pay vehicle tax if you own a car.

The vehicle tax depends on CO2 emission of your car as shown in the table.

Table

Description automatically generated

You want to buy a car for which tax is not more than £30 per year.

Using your graph or the equation of the regression line of F on C, estimate the minimum fuel economy of a car you might buy.

5. Jamie and Lily are investigating different types of correlation

(a) Match each scatter diagram below to the most appropriate type of correlation.

Diagram

Description automatically generated

Jamir and Lily each wear a special bank that measures the number of steps walked in each day (S)

The number of calories burned each day (C).

The tables below show Jamir’s data and Lily’s data for the last eight days.

Table

Description automatically generated

(b) Jamir and Lily want to know if it is justified to use S to estimate C. Calculate the correlation coefficient between S and C for Jamir’s data.

(c) Calculate the product moment for Lily’s data. Hence explain why Jamir’s estimate of X is likely to be more accurate than Lily’s estimate of for any given value S.

(d) Complete the scatter diagram of C against S for Jamir’s data on the grid below. The table with Jamir’s data is repeated below.

Chart, scatter chart

Description automatically generated

(d) Calculate the equation of linear regression of C on S for Jamir’s data.

(e) Jamir wants to burn at least 21000 each week. Work out how many steps he should aim to do each day.