**UNIT-IV**

**Regression and Clustering**

Statistics with R Summarizing data, calculating relative frequencies, tabulating factors and creating contingency tables, testing categorical variables for independence, calculating quantiles of a dataset, converting data into z- scores, t-test, testing sample proportions, testing normality, comparing means of two samples, testing correlation for significance, linear regression in R, logistic regression in R, clustering with R.

**General Statistics**

## Introduction

Any significant application of R includes statistics or models or graphics. This chapter addresses the statistics. Some recipes simply describe how to calculate a statistic, such as relative frequency. Most recipes involve statistical tests or confidence intervals. The statistical tests let you choose between two competing hypotheses; that paradigm is described next. Confidence intervals reflect the likely range of a population parameter and are calculated based on your data sample.

**What is Statistics?**

**Statistics** is an area of applied mathematics concerned with data collection, analysis, interpretation,

and presentation.

This area of mathematics deals with understanding how data can be used to solve complex problems.

Here are a couple of example problems that can be solved by using statistics:

• Your company has created a new drug that may cure cancer. How would you conduct a test

to confirm the drug’s effectiveness?

• You and a friend are at a baseball game, and out of the blue, he offers you a bet that neither

team will hit a home run in that game. Should you take the bet?

• The latest sales data have just come in, and your boss wants you to prepare a report for

management on places where the company could improve its business. What should you look

for? What should you not look for?

**Statistical Terms**

There are various statistical terms that one should be aware of while dealing with statistics.

**Popolation**: A collection of all probable observations of a specific characteristic of interest. It is

a group from which data are collected.

Example: All learners taking data science course.

**Sample**: A subset of population

Example: A group of 20 learners selected for a quiz

**Variable**: An item of interest that can acquire various numerical values

**Categories Of Data**

Data can be categorized into two sub-categories: 1. Qualitative Data 2. Quantitative Data

**Qualitative Data:** Qualitative data deals with characteristics and descriptors that can’t be

easily measured, but can be observed subjectively. Qualitative data is further divided into two

types of data:

**Nominal Data**: Data with no inherent order or ranking such as gender or race.

**Ordinal Data**: Data with an ordered series of information is called ordinal data.

**Quantitative Data:** Quantitative data deals with numbers and things you can measure

objectively. This is further divided into two:

**Discrete Data**: Also known as categorical data, it can hold a finite number of possible values.

Example: Number of students in a class.

**Continuous Data**: Data that can hold an infinite number of possible values. Example: Weight

of a person.

**Types of Statistics**

There are two well-defined types of statistics:

1. Descriptive Statistics

2. Inferential Statistics

**Descriptive Statistics**

Descriptive statistics is a method used to describe and understand the features of a specific data

set by giving short summaries about the sample and measures of the data.

Descriptive Statistics is mainly focused upon the main characteristics of data. It provides a graphical

summary of the data.

Suppose you want to gift all your classmate’s t-shirts. To study the average shirt size of students

in a classroom, in descriptive statistics you would record the shirt size of all students in the class

and then you would find out the maximum, minimum and average shirt size of the class.

Descriptive Statistics is broken down into two categories:

1. Measures of Central Tendency

2. Measures of Variability (spread)

**Measures of Central Tendency**

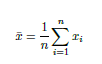
Measures of center are statistics that give us a sense of the “middle” of a numeric variable. Common

measures of center include: - mean - median - mode

**0.2.3 Mean**

Arithmetic average of a range of values or quantities, computed by dividing the total of all values

by the number of values.



**Median** Denotes value or quantity lying at the midpoint of a frequency distribution of observed

values or quantities, such that there is an equal probability of falling above or below it. Simply

put, it is the *middle* value in the list of numbers.



**Measures of Spread**

Measures of spread (dispersion) are statistics that describe how data varies. While measures of

center give us an idea of the typical value, measures of spread give us a sense of how much the data

4

tends to diverge from the typical value. The measures of spread are: - Range - Standard deviation

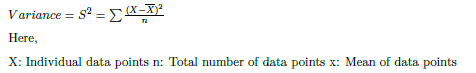
- Variance - Interquartile range

**Range** Range is the difference between the maximum and minimum observations.

*Range* = *max*(*x*) *− min*(*x*)

**Variance** It’s the average distance of the data values from the *mean*. Variance can be calculated

by using the below formula:



**Interquartile Range(IQR)** It is the measure of statistcal dispersion, being equal to the difference

between upper and lower quartile.



Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the

median breaks it in half.

**Standard Deviation** It is the square root of variance. This will have the same units as the data

and mean. It can be calculated by using the below formula:



**Skewness and Kurtosis** Beyond measures of center and spread, descriptive statistics include

measures that give you a sense of the shape of a distribution.

Skewness is a measure of the asymmetry of a data distribution. Skewness is asymmetry in a

statistical distribution, in which the curve appears distorted or skewed either to the left or to the

right. Skewness can be quantified to define the extent to which a distribution differs from a normal

distribution.

Skewness measures the skew while kurtosis measures the “peakedness” of a distribution.

We won’t go into the exact calculations behind skewness and kurtosis, but they are essentially just

statistics that take the idea of variance a step further: while variance involves squaring deviations

from the mean, skewness involves cubing deviations from the mean and kurtosis involves raising

deviations from the mean to the 4th power.

**Inferential Statistics**

Inferential statistics generalize the larger dataset and applies probability theory to draw a conclusion.

It allows you to infer population parameters based on sample statistics and to model relationships

within data.

**Correlation**

A correlation is a statistical test of association between variables that is measured on a -1 to 1

scale. The closer the correlation value is to -1 or 1 the stronger the association, the closer to 0,

the weaker the association. It measures how change in one variable is associated with change in

another variable.

There are a few common types of tests to measure the level of correlation: **Pearson, Spearman,**

**and Kendall**.

Each have their own assumptions about the data that needs to be meet in order for the test to

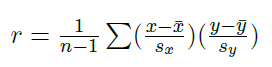
be able to accurately measure the level of correlation. Each type of correlation test is testing the

following hypothesis.

Extent to which two or more variables fluctuate together. A **positive correlation** indicates the

extent to which those variables increase or decrease in parallel; a **negative correlation** indicates

the extent to which one variable increases as the other decreases.



**Pearson correlation assumptions**

Pearson correlation test is a parametric test that makes assumption about the data. In order for

the results of a Pearson correlation test to be valid, the data must meet these assumptions:

• The sample is independently and randomly drawn

• A linear relationship between the two variables is present

**–** When plotted, the lines form a line and is not curved

• There is homogeneity of variance

The variables being used in the correlation test should be continuous and measured either on a

ratio or interval sale, each variable must have equal number of non-missing observations, and there

should be no outliers present.

**Spearman Rank correlation assumptions**

The Spearman rank correlation is a non-parametric test that does not make any assumptions about

the distribution of the data. The assumption for the Spearman rank correlation test is:

• There is a monotonic relationship between the variables being tested

• A monotonic relationship exists when one variable increases so does the other

For the Spearman rank correlation, the data can be used on ranked data, if the data is not normally

distributed, and even if the there is not homogeneity of variance.

**Kendall’s Tau correlation assumptions**

The Kendall’s Tau correlation is a non-parametric test that does not make any assumptions about

the distribution of the data. The only assumption is:

• There should be a monotonic relationship between the variables being tested

The data should be measured on either an ordinal, ratio, or interval scale.

### Null Hypotheses, Alternative Hypotheses, and p-Values

Many of the statistical tests in this chapter use a time-tested paradigm of statistical inference. In the paradigm, we have one or two data samples. We also have two competing hypotheses, either of which could reasonably be true.

One hypothesis, called the null hypothesis, is that nothing happened: the mean was unchanged; the treatment had no effect; you got the expected answer; the model did not improve; and so forth.

The other hypothesis, called the alternative hypothesis, is that something happened: the mean rose; the treatment improved the patients’ health; you got an unexpected answer; the model fit better; and so forth.

We want to determine which hypothesis is more likely in light of the data:

1. To begin, we assume that the null hypothesis is true.
2. We calculate a test statistic. It could be something simple, such as the mean of the sample, or it could be quite complex. The critical requirement is that we must know the statistic’s distribution. We might know the distribution of the sample mean, for example, by invoking the Central Limit Theorem.
3. From the statistic and its distribution we can calculate a p-value, the probability of a test statistic value as extreme or more extreme than the one we observed, while assuming that the null hypothesis is true.
4. If the p-value is too small, we have strong evidence against the null hypothesis. This is called rejecting the null hypothesis.
5. If the p-value is not small, then we have no such evidence. This is called failing to reject the null hypothesis.

There is one necessary decision here: When is a p-value “too small”?

But the real answer is, “it depends.” Your chosen significance level depends on your problem domain. The conventional limit of p < 0.05 works for many problems. In our work, the data are especially noisy and so we are often satisfied with p < 0.10. For someone working in high-risk areas, p < 0.01 or p < 0.001 might be necessary.

In the recipes, we mention which tests include a p-value so that you can compare the p-value against your chosen significance level of α. We worded the recipes to help you interpret the comparison. Here is the wording from Recipe [9.4](https://rc2e.com/generalstatistics#recipe-id125), [Testing Categorical Variables for Independence](https://rc2e.com/generalstatistics#recipe-id125), a test for the independence of two factors:

This is a compact way of saying:

* The null hypothesis is that the variables are independent.
* The alternative hypothesis is that the variables are not independent.
* For α = 0.05, if p < 0.05 then we reject the null hypothesis, giving strong evidence that the variables are not independent; if p > 0.05, we fail to reject the null hypothesis.
* You are free to choose your own α, of course, in which case your decision to reject or fail to reject might be different.

Remember, the recipe states the informal interpretation of the test results, not the rigorous mathematical interpretation. We use colloquial language in the hope that it will guide you toward a practical understanding and application of the test. If the precise semantics of hypothesis testing is critical for your work, we urge you to consult the reference cited under [See Also](https://rc2e.com/generalstatistics#see_also_9) or one of the other fine textbooks on mathematical statistics.

### Confidence Intervals

Hypothesis testing is a well-understood mathematical procedure, but it can be frustrating. First, the semantics is tricky. The test does not reach a definite, useful conclusion. You might get strong evidence against the null hypothesis, but that’s all you’ll get. Second, it does not give you a number, only evidence.

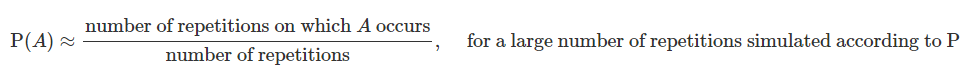
If you want numbers then use confidence intervals, which bound the estimate of a population parameter at a given level of confidence. Recipes in this chapter can calculate confidence intervals for means, medians, and proportions of a population.

For example, Recipe [9.9](https://rc2e.com/generalstatistics#recipe-id123), [“Forming a Confidence Interval for a Mean”](https://rc2e.com/generalstatistics#recipe-id123), calculates a 95% confidence interval for the population mean based on sample data. The interval is 97.16 < μ < 103.98, which means there is a 95% probability that the population’s mean, μ, is between 97.16 and 103.98.

## Summarizing Your Data

## Calculating Relative Frequencies

A simulation involves repeatedly artificially recreating the random phenomenon a large number of times and using the results to investigate properties of interest. In particular, we can use simulation-based relative frequencies to approximate probabilities. That is, the probability of event 𝐴 can be approximated by simulating, according to the assumptions corresponding to the probability measure P, the random phenomenon a large number of times and computing the relative frequency of 𝐴.

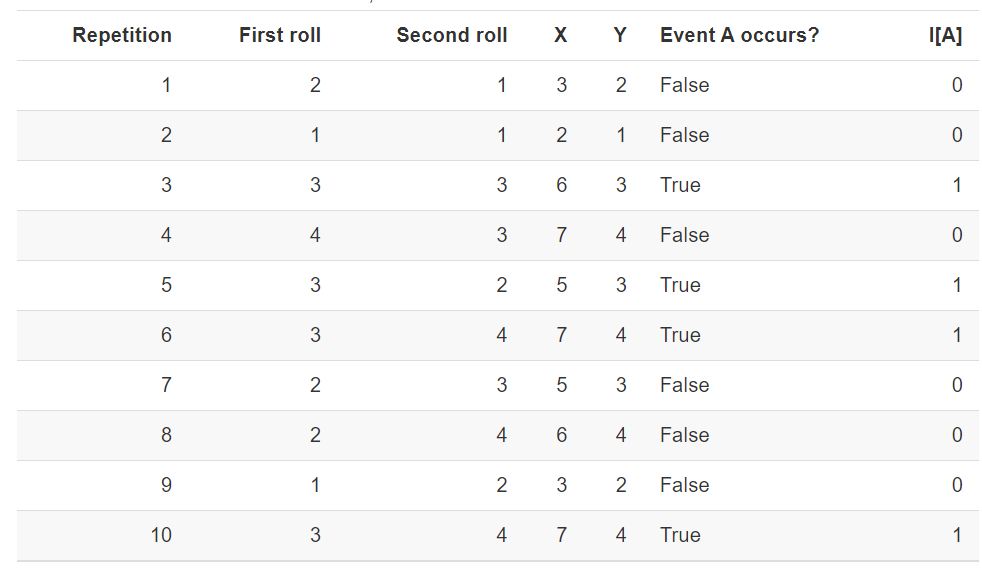


In practice, many repetitions of a simulation are performed on a computer to approximate what happens in the “long run”. However, we often start by carrying out a few repetitions by hand to help make the process more concrete.

**Example 2.29**Use a four-sided die (or a box or a spinner) and perform by hand 10 repetitions of the simulation in Example [2.26](https://bookdown.org/kevin_davisross/probsim-book/introduction-to-simulation.html#exm:dice-sim). (Yes, really do it.) For each repetition, record the results of the first and second rolls (or draws or spins) and the values of X𝑋 and Y𝑌. Based only on the results of your simulation, how would you approximate the following? (Don’t worry if the approximations are any good yet.)

1. P(A)P(𝐴), where A𝐴 is the event that the first roll is 3.
2. P(X=6)P(𝑋=6)
3. P(X≥6)P(𝑋≥6)
4. P(Y=3)P(𝑌=3)
5. P(Y≥3)P(𝑌≥3)
6. P(X=6,Y=3)P(𝑋=6,𝑌=3)
7. P(X≥6,Y≥3)

Results of 10 repetitions of two rolls of a fair four-sided die. X is the sum of the two rolls, Y is the maximum, and A is the event that the first roll is a 3.

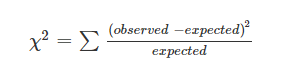


Remember that it is important to distinguish between what entails (1) one repetition of the simulation and its output, and (2) the simulation itself and output from many repetitions. When describing a simulation, refrain from making vague statements like “repeat this” or “do it again”, because “this” or “it” could refer to different elements of the simulation. In the dice example, (1) rolling a die is repeated to generate a single (X,Y)(𝑋,𝑌) pair, and (2) the process of generating (X,Y)(𝑋,𝑌) pairs is repeated to obtain the simulation results. That is, a single repetition involves an ordered pair of die rolls, resulting in an outcome ω𝜔, and the values of the sum X(ω)𝑋(𝜔) and max Y(ω)𝑌(𝜔) are computed for the outcome ω𝜔. The process described in the previous sentence is repeated many times to generate many outcomes and (X,Y)(𝑋,𝑌) pairs according to the probability model.

Think of simulation results being organized in a table like Table, where each row corresponds to a different repetition of the simulation and each column corresponds to a different random variable or event. Remember that indicators are the bridge between events and random variables. On each repetition of the simulation an event either occurs or not. We could record the occurrence of an event as “True/False” for each repetition, or we could record the 1/0 value of the corresponding indicator random variable; see the last two columns in Table for an example.

## Testing Categorical Variables for Independence

Chi-square tests are used to compare relationships between variables measured at the nominal scale. These tests look for differences among frequencies or departures from expected frequencies.



This test has to include its degrees of freedom, which is DF = # categories – 1 – the number of parameters estimated by the data.

The assumptions of this test are:

1. Observations taken at the nominal scale. Categories of the nominal scale are represented as mutually exclusive.
2. Observations are independent of one another.
3. No category has an expected frequency of less than 1, or if there are lots of categories, not more than 20% have an expected frequency < 5.

The test algorithm is probably familiar for people who have taken genetics:

o test the independence of two categorical variables, we commonly use the Chi-Square Test of Independence. Here’s a step-by-step guide on how to perform this test:

1. Formulate Hypotheses

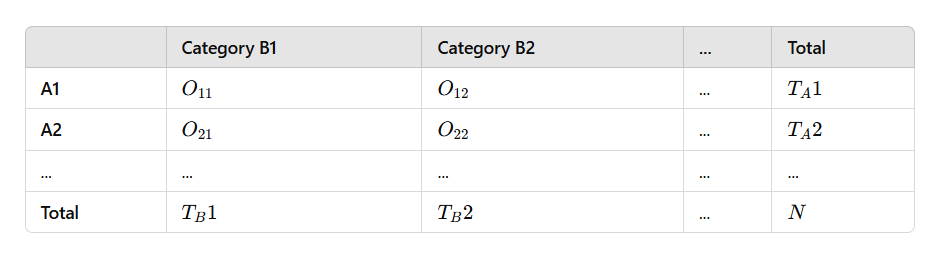
•Null Hypothesis (𝐻0H0): The two categorical variables are independent.

•Alternative Hypothesis (𝐻𝑎Ha): The two categorical variables are not independent.

2. Collect and Organize Data

Create a contingency table showing the frequencies of the different categories for the two variables.

Example:



Where 𝑂𝑖𝑗Oij is the observed frequency for the combination of the 𝑖i-th category of variable A and the 𝑗j-th category of variable B.

3. Calculate Expected Frequencies

The expected frequency for each cell, 𝐸𝑖𝑗Eij, is calculated under the assumption of independence: 𝐸𝑖𝑗=𝑇𝐴𝑖×𝑇𝐵𝑗𝑁Eij=NTAi×TBj

4. Compute the Chi-Square Statistic

The Chi-Square statistic is calculated as: 𝜒2=∑(𝑂𝑖𝑗−𝐸𝑖𝑗)2𝐸𝑖𝑗χ2=∑Eij(Oij−Eij)2 Where the summation is over all cells in the contingency table.

5. Determine Degrees of Freedom

The degrees of freedom (df) for the test is given by: 𝑑𝑓=(𝑟−1)×(𝑐−1)df=(r−1)×(c−1) Where 𝑟r is the number of rows (categories of variable A) and 𝑐c is the number of columns (categories of variable B).

6. Compare with the Critical Value

Compare the calculated 𝜒2χ2 statistic with the critical value from the Chi-Square distribution table at the desired significance level (𝛼α) and the calculated degrees of freedom.

7. Make a Decision

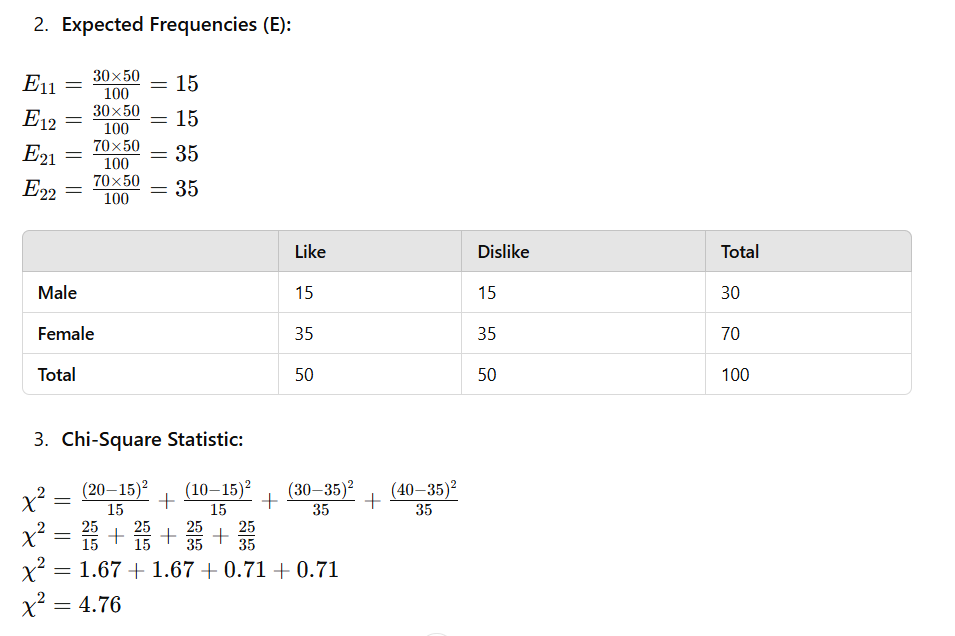
•If 𝜒2χ2 calculated > 𝜒2χ2 critical, reject the null hypothesis (𝐻0H0).

•If 𝜒2χ2 calculated ≤ 𝜒2χ2 critical, do not reject the null hypothesis.

Example

Consider a study examining the relationship between gender (Male, Female) and preference for a new product (Like, Dislike).

## 



## 

## Converting Data to z-Scores

### Problem

You have a dataset, and you want to calculate the corresponding z-scores for all data elements. (This is sometimes called normalizing the data.)

### Solution

Use the scale function:

**scale**(x)

*#> [,1]*

*#> [1,] 2.2638*

*#> [2,] -0.0463*

*#> [3,] -0.9400*

*#> [4,] -0.7755*

*#> [5,] 1.0020*

*#> [6,] 0.0650*

*#> [7,] -1.0227*

*#> [8,] -0.5412*

*#> [9,] 0.1408*

*#> [10,] -0.1458*

*#> attr(,"scaled:center")*

*#> [1] 2.53*

*#> attr(,"scaled:scale")*

*#> [1] 2.15*

This works for vectors, matrices, and data frames. In the case of a vector, scale returns the vector of normalized values. In the case of matrices and data frames, scale normalizes each column independently and returns columns of normalized values in a matrix.

### Discussion

You might also want to normalize a single value y relative to a dataset x. You can do so by using vectorized operations as follows:

(y - **mean**(x)) / **sd**(x)

*#> [1] -0.619*

## Testing the Mean of a Sample (t Test)

## Forming a Confidence Interval for a Mean

## Forming a Confidence Interval for a Median

## Testing a Sample Proportion

## Forming a Confidence Interval for a Proportion

## Testing for Runs

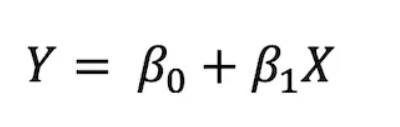
## Comparing the Means of Two Samples

## Testing a Correlation for Significance

LINEAR REGRESSION

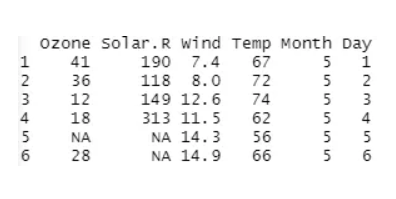
## **Simple Linear Regression**

Simple linear regression, or univariate linear regression model is one type of linear regression model that has a single independent variable. This can be expressed as the equation below where X represents one independent variable and Y represents the dependent variable.



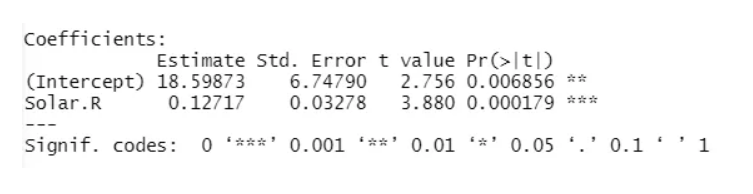
For linear regression, we use **lm()** function in R which allows us to fit linear models. Let us use airquality dataset from R which contains daily air quality measurements in New York that are measured from May to September in 1973.

head(airquality)



Let’s test, with a linear regression model, how the level of solar radiation(Solar.R) is related to mean ozone concentration(Ozone).

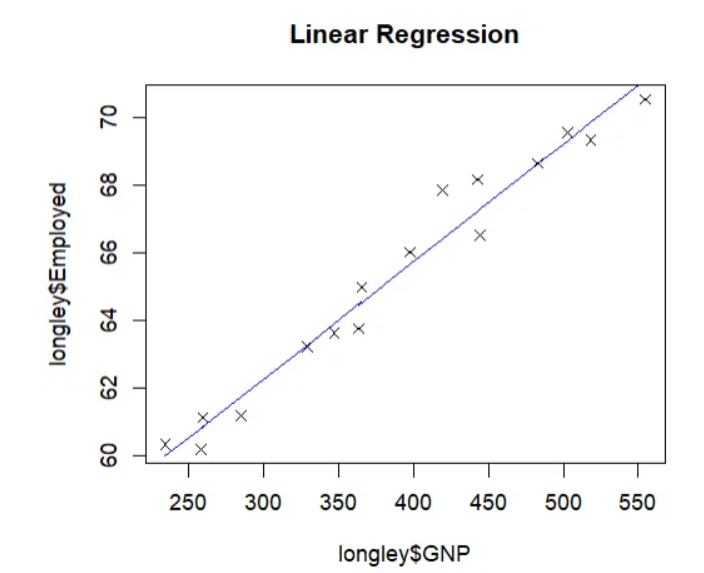
model1 <- lm(Ozone ~ Solar.R, data = airquality)  
summary(model1)



In R, the dependent variable and independent variables are separated by ‘~’ symbol. In our model, the dependent variable is ‘Ozone’ and the independent variable is ‘Solar.R’. Since the p-value for our independent variable ‘Solar.R’ is significant, we can say that the relationship between the ozone concentration and solar radiation is significant. It is always important to check the **p-value** first because if it is not significant, the model is invalid in the end.

Let us now graph the relationship between the solar radiation and ozone concentration.

plot(longley$GNP, longley$Employed, pch = 1, bg = 'cyan', main = 'Linear Regression')  
lines(longley$GNP, model1$fitted.values, col = 'red')



If we want to go further to predict a certain dependent variable value with a given independent variable value, we can use the function **predict()**.

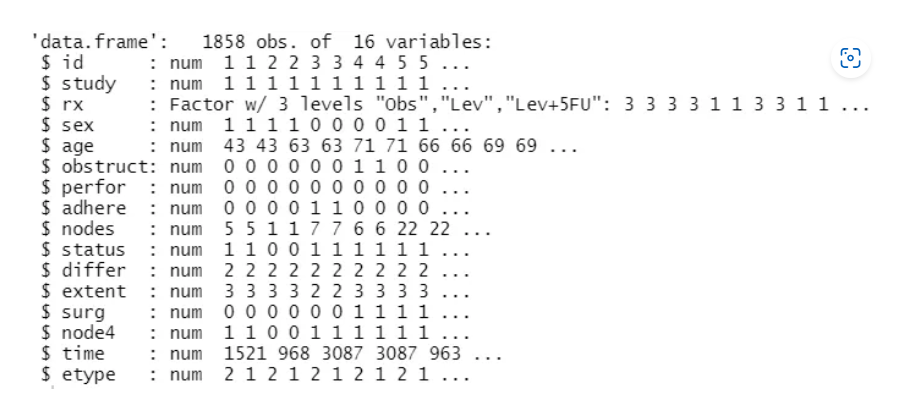
predict(model1, list(GNP = c(400)))

Running the command above returns 65.74451, which is the predicted value of employed rate when the GNP is 400.

Logistic Regression

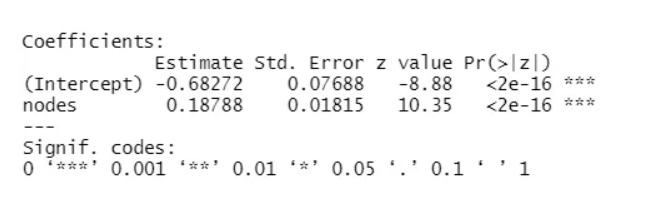
Logistic regression, as mentioned before, is used for solving classification problems. Let’s use a different dataset called colon from the library ‘survival’ to understand how logistic regression works with R.

library(survival)  
str(colon)



This dataset shows survival and recurrence information on 929 people from a clinical trial on colon cancer chemotherapy. Our dependent variable is ‘status’ which takes 1 for colon cancer recurrence or death, and 0 otherwise. For simple logistic regression, we will chose the variable ‘nodes’, the number of lymph nodes with cancer cell.

colon <- na.omit(colon)  
model3 <- glm(status ~ nodes, family = binomial(logit), data = colon)  
summary(model3)



As can be seen, the p-value for ‘nodes’ is significant and the coefficient is positive.