Syntax:

glm (formula, data,family)

- The symbol expressing the relationship between the variables is a formula.
- The data set containing the values of these variables is known as data.
- family is a R object that specifies the model's details. For logistic regression, it has a binomial value.

Input Data: Let's take the R inbuilt data set "mtcars", which provides details of various car models & engine specifications. The transmission mode of the car i.e. whether the car is manual or automatic is described by the column *am* having a binary value as 0 or 1. You can create the model between columns "am" (Outcome/ dependent/ response variable) and three others – hp, wt and cyl (predictor variables).This model is aimed at determining, if car would have manual or automatic transmission, given the horse power (hp), weight (wt) and number of cylinders (cyl) in the car.

```
> input <- mtcars[,c("am","cyl","hp","wt")]
>
> print(head(input))
                  am cyl  hp    wt
Mazda RX4          1   6 110 2.620
Mazda RX4 Wag      1   6 110 2.875
Datsun 710         1   4  93 2.320
Hornet 4 Drive     0   6 110 3.215
Hornet Sportabout  0   8 175 3.440
Valiant            0   6 105 3.460
>
```

Figure 15.12: The sample data set for logistic regression

```
> am.data = glm(formula = am ~ cyl + hp + wt, data = input, family = binomial)
> print(summary(am.data))

Call:
glm(formula = am ~ cyl + hp + wt, family = binomial, data = input)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.17272  -0.14907  -0.01464   0.14116   1.27641

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 19.70288    8.11637   2.428   0.0152 *
cyl          0.48760    1.07162   0.455   0.6491
hp           0.03259    0.01886   1.728   0.0840 .
wt          -9.14947    4.15332  -2.203   0.0276 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.2297  on 31  degrees of freedom
Residual deviance:  9.8415  on 28  degrees of freedom
AIC: 17.841

Number of Fisher Scoring iterations: 8
```

Figure 15.13: The logistic regression model

The null deviance demonstrates how well a model with an intercept term can predict the dependent variable, whereas the residual deviance represents how well a model with n predictor variables can predict the dependent variable. Deviance is measure of goodness of fit of a model.

In the summary as the p-value is more than 0.05 for the variables "cyl" (0.0152) and "hp" (0.0276), we will consider them insignificant in contributing to the value of the variable "am". Only weight (wt) impacts the "am" value in this regression model.

15.7 TIME SERIES ANALYSIS

A Time Series is any metric that is measured at regular intervals. It entails deriving hidden insights from time-based data (years, days, hours, minutes) in order to make informed decisions. When you have serially associated data, time series models are particularly beneficial. Weather data, stock prices, industry projections, and so on are just a few examples.

A time series is represented as follows:

A data point, say (Y_t), at a specific time t (indicated by subscript t) is defined as the either sum or product of the following three components:

Seasonality (S_t), Trend (T_t); and Error (e_t) (also known as, **White Noise**).

Input: Import the data set and then use ts() function.
The steps to use the function are given below. However, it is pertinent to note here that the input values used in this case should ideally be a numeric vector belonging to the "numeric" or "integer" class.
The following functions will generate quarterly data series from 1959:
ts(inputData, frequency =4, start = c(1959,2)) #frequency 4 => QuarterlyData
The following function will generate monthly data series from 1990
ts(1:10, frequency =12, start = 1990) #freq 12 => MonthlyData
The following function will generate yearly data series from 2009 to 2014.
ts(inputData, start=c(2009), end=c(2014), frequency=1) # YearlyData

In case, you want to use Additive Time Series, you use the following:

$$Y_t = S_t + T_t + e_t$$

However, for Multiplicative Time Series, you may use:

$$Y_t = S_t \times T_t \times e_t$$

The additive time series can be converted from multiplicative time series by taking using the log function on the time series as represented below:

$$additiveTS = log(multiplcativeTS)$$

15.7.1 Stationary Time Series

A time series is considered "stationary" if the following criteria are satisfied:

1. When the mean value of a time series remains constant over a period of time and hence, the trend component is removed Over time, the variance does not increase.
2. Seasonality has a minor impact.

This means it has no trend or seasonal characteristics, making it appear to be random white noise regardless of the time span viewed.

Steps to convert a time series as stationary
Each data point in a time series is differentiated by subtracting it from the one before it. It is a frequent technique for making a time series immobile. To make a stationary series out of most time series patterns 1 or 2 differencing is required.

15.7.2 Extraction of trend, seasonality and error

Using decompose() and forecast::stl, the time series is separated into seasonality, trend, and error components (). You may use the following set of commands to do so.

```
timeSeriesData = EuStockMarkets[,1]
resultofDecompose = decompose(timeSeriesData, type="mult")
plot(resultofDecompose)
resultsofSt1 = stl(timeSeriesData, s.window = "periodic")
```

15.7.3 Creating lags of a time-series

A lag of time series is generated when the time basis is shifted by a given number of periods. Moreover, the state of a time series a few periods ago, however, may still have an effect on its current state. Hence, in the time series models, the delays of a time series are typically used as explanatory variables.

```
lagTimeSeries = lag(timeSeriesData, 3) #Shifting to 3 periods earlier
library(DataCombine)
mydf = as.data.frame(timeSeriesData)
mydf = slide(mydf, "x", NewVar = "xLag1", slideBy = -1) #create lag1
variable
mydf = slide(mydf, "x", NewVar = "xLag1", slideBy = 1)
```

Check your Progress 2

1. What is logistic regression?

   ……..

2. What are the uses of Time-Series analysis?

   …….

3. Differentiate between linear regression and logistic regression?

   ……

15.8 SUMMARY

This unit introduces the concept of data analysis and examine its application using R programming. It explains about the Chi-Square Test that is used to determine if two categorical variables are significantly correlated and further study its application on R. The unit explains the Regression Analysis, which is a common statistical technique for establishing a relationship model between two variables- a predictor variable and the response variable. It further explains the various models in Regression Analysis including Linear and Logistics Regression Analysis. In Linear Regression the two variables are related through an equation of degree is one and employs a straight line to explain the relationship between variables. It is categorised into two types- Simple Linear Regression which uses only one independent variable and Multiple Linear Regression which uses two or more independent variables. Once familiar with the Regression, the unit proceeds to explain about the logistic regression, which is a classification algorithm for determining the probability of event success and failure. It is also known as Binomial logistic regression and is based on the sigmoid function, with probability as the output and input ranging from $-\infty$ to $+\infty$. At the end, the unit introduces the concept of time series analysis and help understand its application and usage on R. It also discusses the special case of Stationary Time Series and how to make a time series stationary. This section further explains how to extract the trend, seasonality and error in a time series in R and the creating lags of a time series.

15.9 ANSWERS

Check your Progress 1

1. A regression model that employs a straight line to explain the relationship between variables is known as linear regression. In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is one. It searches for the value of the regression coefficient(s) that minimises the total error of the model to find the line of best fit through your data.
2. The Chi-square test of independence determines whether there is a statistically significant relationship between categorical variables. It's a hypothesis test that answers the question—do the values of one categorical variable depend on the value of other categorical variables?
3. Linear regression considers 2 variables whereas multiple regression consists of 2 or more variables.

Check your Progress 2

1. Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring.
2. Time series analysis is used to identify the fluctuation in economics and business. It helps in the evaluation of current achievements. Time series is used in pattern recognition, signal processing, weather forecasting and earthquake prediction.
3. The problems pertaining to regression are solved using linear regression; however, the problems pertaining to classification are solved using the logistic regression. The linear regression yields a continuous result, whereas logistic regression yields discrete results.

15.10 REFERENCES AND FURTHER READINGS

1. De Vries, A., & Meys, J. (2015). *R for Dummies*. John Wiley & Sons.
2. Peng, R. D. (2016). *R programming for data science* (pp. 86-181). Victoria, BC, Canada: Leanpub.
3. Schmuller, J. (2017). *Statistical Analysis with R For Dummies*. John Wiley & Sons.
4. Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.
5. Lander, J. P. (2014). *R for everyone: Advanced analytics and graphics*. Pearson Education.
6. Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.
7. Heumann, C., & Schomaker, M. (2016). *Introduction to statistics and data analysis*. Springer International Publishing Switzerland.
8. Davies, T. M. (2016). *The book of R: a first course in programming and statistics*. No Starch Press.
9. https://www.tutorialspoint.com/r/index.html
10. https://data-flair.training/blogs/chi-square-test-in-r/
11. http://r-statistics.co/Time-Series-Analysis-With-R.html
12. http://r-statistics.co/Logistic-Regression-With-R.html