

Business Analytics in R

Introduction to Statistical Programming

TIMOTHY WONG¹
SECOND AUTHOR²

March 12, 2018

¹timothy.wong@hotmail.co.uk

²person@organisation.com

Contents

1	Basics	3
2	Regression Models	5
2.1	Linear Regression	5
2.2	Poisson Regression	14
2.3	Logistic Regression	16
3	Tree-based Methods	19
3.1	Decision Trees	20
3.2	Random Forest	22
4	Neural Networks	25
4.1	Multilayer Perceptron	27
5	Time Series Analysis	33
5.1	Auto-Correlation Function	33
5.2	Decomposition	38
5.3	ARIMA Model	41
6	Survival Analysis	47

6.1	Kaplan-Meier Estimator	47
6.2	Cox Proportional Harzard Model	49
7	Unsupervised Learning	53
7.1	K -means Clustering	54
7.2	Hierarchical Clustering	56

List of Figures

2.1	Regression diagnostic plots	11
2.2	A less flexible model showing better generalisability	13
2.3	A more flexible model illustrating the risk of overfitting	14
2.4	Poisson distribution with different λ values	15
2.5	Graph showing the range of a logistic function	16
3.1	Recursive partitioning	19
3.2	Decision tree for a regression problem	21
4.1	Common neural activation functions	26
4.2	MLP model with two hidden layers	29
4.3	MLP model with one-hot encoded categorical variables	31
5.1	ACF and PACF correlograms	37
5.2	Lag plot showing the correlation of various lag periods	38
5.3	Decomposing an additive time series	39
5.4	Linear time series forecasting with trend and seasonal components	40
5.5	Forecast generated from a seasonal ARIMA model.	44
6.1	Kaplan-Meier curves showing two strata	48

6.2	Scaled Schoenfeld residuals of a selected set of variables plotted against time	51
7.1	Different ways to cluster a set of unlabelled objects	53
7.2	Biplot showing the first and second principal components	55
7.3	Comparing K -means clustering results using different K values	56
7.4	Iterative steps of agglomerative hierarchical clustering	59
7.5	Dendrogram illustrating hierarchical clustering using complete linkage	60
7.6	Fan phylogram showing hierarchical clusters	62

List of Tables

2.1	Summary of key model information	10
5.1	Description of the Amprion dataset	34
5.2	Description of the Breman weather dataset	35

Preface

Lorem ipsum dolor sit amet.

“ *Jim Hacker: The statistics are irrefutable...*
Sir Humphrey: Statistics? You can prove anything with statistics.
Jim Hacker: Even the truth.
Sir Humphrey: Yes... No! ”

Yes, Prime Minister, 1986

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Structure of book

Each unit will focus on <SOMETHING>.

About the companion website

The website for this file contains:

- A link to (freely downloadable) latest version of this document.

- [Link to download LaTeX source for this document.](#)
- Miscellaneous material (e.g. suggested readings etc).

Acknowledgements

- A special word of thanks goes to.
- I'll also like to thank.
- I'm deeply indebted my parents, colleagues and friends for their support and encouragement.

Tim
timothy.wong@hotmail.co.uk

Chapter 1

Basics

Chapter 2

Regression Models

Within the realm of supervised learning, regression model is one of the most widely used statistical techniques. It provides a way to quantify the relationship between dependent variable Y and independent variable X , where $Y \subseteq \mathbb{R}$ and $X = \{x_1, x_2, \dots, x_M\}$.

2.1 Linear Regression

Linear regression is a very popular parametric statistical technique. This means it assumes independent variables are normally distributed (i.e. having a bell-shaped curve) and the model residuals also follow the same distribution. The goal of linear regression is to minimise the sum of squared residual. Regression method is extremely useful in the sense that the output can quantify the effects of each input variable, it also informs users whether the effects are statistical significant or not.

In a univariate scenario where x is the only input, it provides the best fit for the model equation $\hat{y} = \beta_0 + \beta_1 x$. The parameter β_0 is normally referred as the intercept, while the β_1 value is called the slope. This is illustrated in the chart on the left side in the figure below.

Simple linear model can be extended to include a high-order polynomial term of variable x . It provides higher model flexibility so that linear model fits the data better. For example, an M^{th} order polynomial term has been fitted to the dataset on the next chart. The choice of M is subjective but usually a small value of M is desirable because it helps avoid overfitting.

In R language you can simply call the function `lm` to perform linear regression. You need to provide at least two arguments (`data` and `formula`), extra arguments

can be provided such as `subset`. For example, `lm(y ~ x, myData)` will perform a simple univariate linear model using independent variable x to predict dependent variable y in the data frame object `myData`.

Dummy variable can be easily created from categorical labels. Users can simply use the syntax `lm(y ~ x1 + x2, myData)` where `x1` is a numeric variable and `x2` either a factor or a character variable. The `lm` function will automatically create dummy variables on-the-fly. Normally the first category in the column will be used as reference level. If users would like to control the reference item, `x2` can be coded as a factor variable with levels defined.

A more flexible regression model can be built by extending a linear model with polynomial terms. It can be expressed as `lm(y ~ poly(x1, 3) + x2, myData)`. In this case, we are defining a cubic relationship with variable `x1` and linear relationship with variable `x2` (Five coefficients in total will be estimated, four from the regression terms plus one from the intercept). Such model can be expressed mathematically as equation (4.2b).

$$\hat{y} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3}_{\text{Cubic polynomial term}} + \underbrace{\beta_4 x_2}_{\text{Linear term}} \quad (2.1)$$

Interaction refers to the combined effect (also known as synergy) of more than one independent variables. For example, independent variable `x1` and `x2` might have no effect on dependent variable y alone. However, the effect on y can become prominent when these two variables get combined. In R language you can use the syntax `lm(y ~ x1 * x2)` to represent the relationship. The function `I(x1 * x2)` can be used to suppress interaction term and the arguments will be treated as simple arithmetic operations.

Exercise 1 Simple Linear Regression

In this exercise, we are going to predict car efficiency using the `mtcars` teaching dataset. The dataset is embedded in open source R and can be called directly by `mtcars`. You can browse the top several rows by executing `head(mtcars)` in your R console. You can read the definition of individual columns by calling the command `?mtcars`.

```
head(mtcars)
```

```
##           mpg  cyl  disp  hp  drat    wt   qsec  vs  am  gear  carb
## Mazda RX4      21.0   6   160  110 3.90 2.620 16.46  0   1     4     4
## Mazda RX4 Wag  21.0   6   160  110 3.90 2.875 17.02  0   1     4     4
## Datsun 710      22.8   4   108   93 3.85 2.320 18.61  1   1     4     1
## Hornet 4 Drive  21.4   6   258  110 3.08 3.215 19.44  1   0     3     1
## Hornet Sportabout 18.7   8   360  175 3.15 3.440 17.02  0   0     3     2
## Valiant        18.1   6   225  105 2.76 3.460 20.22  1   0     3     1
```

Before running any models, we can explore the dataset a bit further by visualising it. The R package `ggplot2` is a very popular visualisation add-in. It builds charts by stacking layers of graphics on it. You may try the following lines:

```

# Create a simple histogram using base R plot
hist(mtcars$mpg)
# Using dplyr pipeline style code (equivalent output)
library(dplyr)
mtcars$mpg %>% hist()
# Histogram using ggplot package
library(ggplot2)
mtcars %>%
  ggplot(aes(x=mpg)) +
  geom_histogram() +
  labs(x='Miles-per-gallon',
       y='Count',
       title='Histogram showing the distribution of car performance')
# Scatterplot showing the relationship between mpg and wt
library(ggplot2)
mtcars %>%
  ggplot(aes(x=wt, y=mpg, colour=factor(cyl))) +
  geom_point() +
  labs(x='Weight',
       y='Miles-per-gallon',
       colour='Number of cylinders',
       title='Scatterplot showing car weight against performance')
# Create a boxplot showing mpg distribution of different gear types
mtcars %>%
  ggplot(aes(x=factor(am,
                    levels = c(0,1),
                    labels = c('Automatic', 'Manual')), y=mpg)) +
  geom_boxplot() +
  labs(x='Gear type',
       y='Miles-per-gallon')
# Draw a scatterplot with facets to visualise multiple variables
mtcars %>%
  ggplot(aes(x=hp, y=mpg, colour=factor(gear), size=disp)) +
  geom_point() +
  facet_grid(. ~ cyl) +
  labs(x='Horsepower',
       y='Miles-per-gallon',
       colour='Number of gears',
       size='Displacement')
# Draw a matrix scatterplot
pairs(mtcars)
# Load the package GGally and use the function ggpairs
# Draws a prettier matrix scatterplot
library(GGally)
ggpairs(mtcars)

```

Now let us try to model car efficiency using the mpg column as dependent variable. The hypothesis is that we suspect heavy cars have lower miles-per-gallon. A simple way to investigate this is to build a univariate linear model using the `lm` function and analyse the results. The following code would fit an ordinary least-squares (OLS) regression model using one predictor variable. The function `summary(myModel1)` would print out all the crucial results of the model.


```
# Build a univariate linear model
myModel1 <- lm(mpg ~ wt, mtcars)
summary(myModel1)

##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776   19.858 < 2e-16 ***
## wt          -5.3445     0.5591   -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10
```

You can also build more complex multivariate models using the same command. You can instruct the function `lm()` to deal with nominal and ordinal variables by explicitly stating `factor(myVar)` in the formula. Alternatively, if your variable is already in factor data type, the regression function would handle it automatically without stating it in the formula.

```
# To build a multivariate linear model with polynomial terms.
# Dummy variables are automatically created on-the-fly.
myModel2 <- lm(mpg ~ factor(cyl) + disp + poly(hp, 3) +
              drat + qsec +
              factor(am) + factor(gear), mtcars)
summary(myModel2)
# Multivariate model with interaction effect
myModel3 <- lm(mpg ~ factor(cyl) * disp + poly(hp, 3) +
              drat + qsec +
              factor(am) + factor(gear), mtcars)
summary(myModel3)
```

If you manage to produce the model summary, you will find a lot of useful information there:

Table 2.1: Summary of key model information

Term	Description
Residuals	This is the unexplained bit of the model, defined as observed value minus fitted value ($\epsilon = y_i - \hat{y}_i$). If the model's parametric assumption is correct, the mean and median values of the residuals should be very close to zero. The distribution of the residuals should have equal tails on both ends.
Estimate	Coefficient of the corresponding independent variable (i.e. the β values).
Standard error	Standard deviation of the estimate.
t -value	The number of standard deviations away from zero (i.e. the null hypothesis).
$P(> t)$	p -value of the model estimate. In general, variables with p -value above 0.05 are considered statistical significant.
Multiple R^2	Pearson's correlation squared which indicates strength of relationship between the observed and fitted values.
Adjusted R^2	Adjusted version of R^2 .
F -statistic	Global hypothesis for the model as a whole.

As the input predictor variables have different scale, the model coefficient are not directly comparable yet. To compare the effect magnitude of the predictor variables, they need to be standardised first. The `QuantPsyc` package has a function `lm.beta()` which performs coefficient standardisation. The magnitude of the standardised coefficient indicates the relative influence of each predictor variable.

```
library(QuantPsyc)
lm.beta(myModel1)
lm.beta(myModel2)
lm.beta(myModel3)
```

Users can also compare multiple nested models¹ using ANOVA technique and check for any statistically significant difference. The code snippet below shows an example.

```
# Compare all models using Chi-square test
anova(myModel1, myModel2, myModel3, test='Chisq')
```

¹Refers to models having additional predictor terms. For example, $\hat{y} = x_1 + x_2 + x_3$ and $\hat{y} = x_1 + x_2 + x_3 + x_4$ are both nested models of $\hat{y} = x_1 + x_2$.

Exercise 2 Regression Diagnostics

As we have discussed previously, linear regression is a parametric method which means it has strong underlying assumptions. Studying the model's diagnostic measurements would help us understand whether the assumptions are sufficiently met. You may use the `plot()` function to create a series of regression diagnostic plots. The following command generates several diagnostic plots for a standard OLS regression model:

```
plot(myModel3)
```

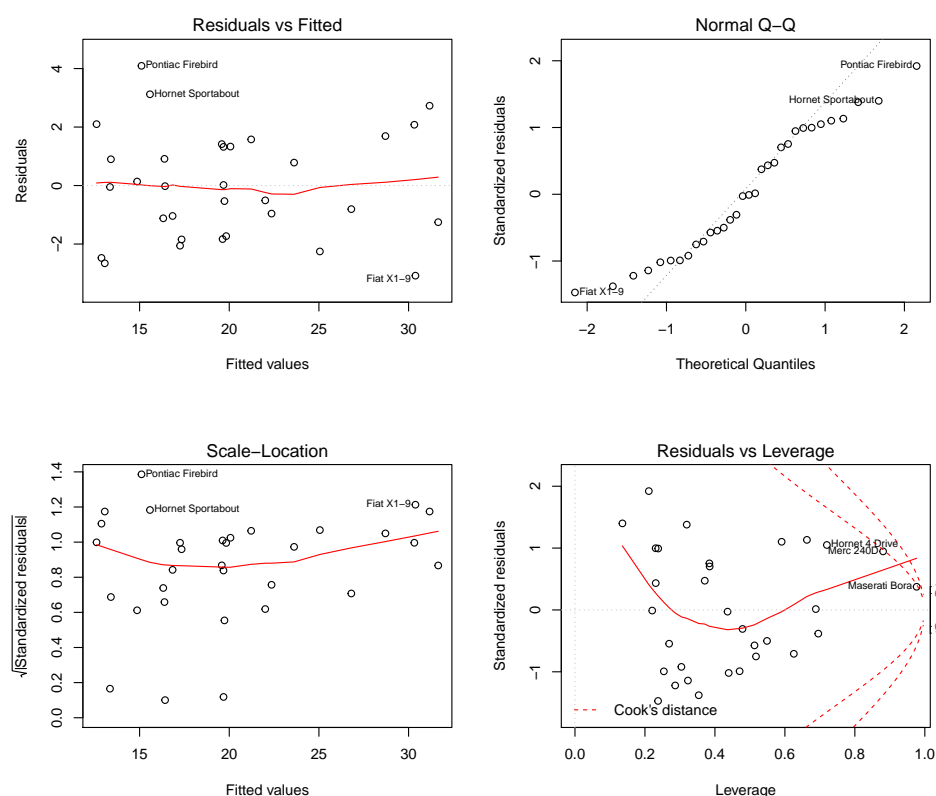


Figure 2.1: Regression diagnostic plots

- **Residuals vs Fitted** Checks for non-linear relationship. For a good linear fit, you should find the residuals evenly distributed along a straight line centred near zero (i.e. a horizontal line). Contrarily, curvy line indicate poor model fit with possible non-linear effects not yet captured.
- **Normal Quantile-Quantile** One the of main assumptions of OLS regression is that the residual is drawn from zero-centred Gaussian distribution $\epsilon_i \sim$

$\mathcal{N}(0, \sigma^2)$. To verify whether the proposed model satisfy the assumptions, we can use a normal quantile-quantile plot (Q-Q plot) to perform a quick check. The plot aligns model residuals against a theoretical normal distribution. If the residuals spread along a straight line on the Q-Q plot, it suggests that the residuals are normally distributed. Alternatively, if the data points deviate from the line it indicates vice versa. In this case, parametric model assumption does not hold and you might have to consider improving your model.

- **Scale-Location** Shows the distribution of the standardised residuals along the range of fitted values. As standard OLS model is assumed to be homoscedastic, the residual variance is not expected to vary along the range of fitted values (i.e. expects a near-horizontal line). If the standardised residual forms a distinguishable pattern (e.g. fanning out), then the model may be heteroscedastic and hence violates the underlying assumptions.
- **Residual vs Leverage (Cook's Distance)** This plot indicates the observation which are highly influential to the model. For observations having high leverage, they pose influence to the model. This means that the model estimates are strongly affected by these cases. If such observations have high residuals (i.e. large Cook's distance), they can sometimes be considered as outliers. On the other hand, most observations would have low leverage and short Cook's distance. This means that the model estimates would not have varied a lot if few such observations were to be added or discarded.

Exercise 3 Model Overfitting

Linear regression can be made more flexible by increasing the order of the polynomial terms. This allows linear model to capture non-linear effects. However, one of the major problems of a flexible model is that it risks overfitting the data. This means that the model might appear to have very good fit during training, but it may fit poorly when it is tested using unseen data. In general, overfitted models have very little inference and are ungeneralisable.

The following code snippet runs an OLS model with variable level of flexibility.

```

# Bivariate linear model with polynomial terms
# You can change the values here.
J <- 3
K <- 2
myModel4 <- lm(mpg ~ poly(wt,J) + poly(hp,K), mtcars)
summary(myModel4)
# Create the base axes as continuous values
wt_along <- seq(min(mtcars$wt), max(mtcars$wt), length.out = 50)
hp_along <- seq(min(mtcars$hp), max(mtcars$hp), length.out = 50)
# Use the outer product of wt and hp to run predictions for every
# point on the plane
f <- function(k1, k2, model){ z <- predict(model, data.frame(wt=k1, hp=k2 )) }
myPrediction <- outer(wt_along, hp_along, f, model = myModel4)
# Draw the 3D plane
myPlane <- persp(x = wt_along, xlab = 'Weight',
                 y = hp_along, ylab = 'Horsepower',
                 z = myPrediction, zlab = 'Miles-per-Gallon',
                 main = 'Fitted vs observed values in 3D space',
                 theta = 30, phi = 30, expand = 0.5, col = "lightblue")
# Add original data as red dots on the plane
myPoints <- trans3d(x = mtcars$wt,
                   y = mtcars$hp,
                   z = mtcars$mpg,
                   pmat=myPlane)
points(myPoints, col='red')

```

$$\hat{y} = \beta_0 + \sum_{j=1}^3 \beta_{wtj} x_{wt}^j + \sum_{k=1}^2 \beta_{hp_k} x_{hp}^k$$

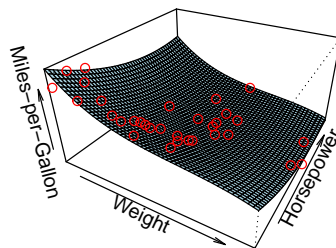


Figure 2.2: A less flexible model showing better generalisability

$$\hat{y} = \beta_0 + \sum_{j=1}^8 \beta_{wtj} x_{wt}^j + \sum_{k=1}^5 \beta_{hp_k} x_{hp}^k$$

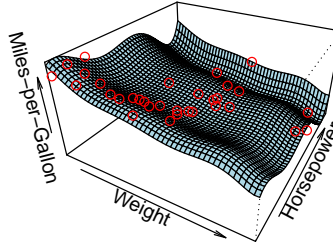


Figure 2.3: A more flexible model illustrating the risk of overfitting

2.2 Poisson Regression

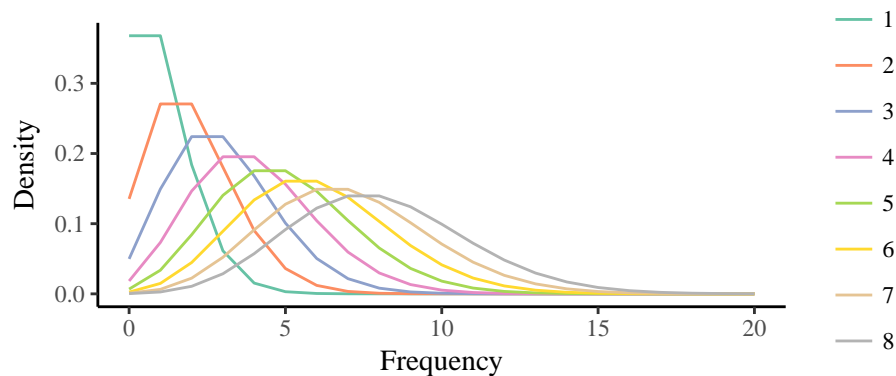
In the previous section, simple OLS model assumes the dependent variable follows a Gaussian distribution which spans the range $(-\infty, +\infty)$. Yet sometimes we would like to estimate the discrete number of events where it is often a positive integer ($\mathbb{N} = \{0, 1, 2, 3, \dots\}$). In this case, Poisson distribution can be used. Poisson distributions can be found in everyday life, the following are typical examples:

- Number of children in a household.
- Number of bank notes in a wallet.

Poisson regression model assumes the response variable is drawn from Poisson distribution². It can take into account multiple predictor variables. The Poisson regression equation is defined in the equations below (2.2).

$$\hat{y} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_M x_M} \quad \ln \hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_M x_M \quad (2.2)$$

²A Poisson distribution is defined by a single parameter $y \sim \text{Poisson}(\lambda)$, where $\sigma^2 = \lambda$ and $\mu = \lambda$. This means that for every Poisson distribution, the mean μ and variance σ^2 must be equal.

Figure 2.4: Poisson distribution with different λ values

Exercise 4 Testing for Poisson Distribution

Using the `mtcars` dataset, we can estimate the number of carburetors (i.e. the variable `carb`) in cars using predictor variables available. You can use the command `hist(mtcars$carb)` to draw a simple histogram. You should find that this variable (1) never goes below zero and (2) has a long but thin tail towards the positive side.

```
# Draw a simple histogram.
hist(mtcars$carb)
# Compute the mean and variance.
mean(mtcars$carb)
var(mtcars$carb)
```

To robustly check whether a variable is truly drawn from a Poisson distribution, one of the most common methods is to perform a Chi-squared goodness-of-fit test. If the resulting p -value is small enough, we can then accept the variable as Poisson.

```
# Performs the Chi-squared goodness-of-fit test.
# It checks whether the variable is drawn from a Poisson distribution.
library(vcd)
gf <- goodfit(mtcars$carb, type= "poisson", method= "ML")
# Plots the observed frequency vs fitted Poisson distribution.
# The hanging bars should fill the space if it was perfectly Poisson.
plot(gf)
# Checks the statistical p-value of the goodness-of-fit test.
# If p<=0.05 then it is safe to say that the variable is Poisson.
summary(gf)

##
## Goodness-of-fit test for poisson distribution
##
##               X^2 df      P(> X^2)
## Likelihood Ratio 20.53973  4 0.0003906369
```

Exercise 5 Building a Poisson Model

The following lines produce a Poisson regression model using the `mtcars` dataset. The variable `carb` is used as dependent variable while `hp`, `wt` and `am` are used as independent predictor variables. The regression output can be interpreted in a similar way as the ones from linear model.

```
# Build a Poisson model to predict the number of carburetors in a car.
myModel5 <- glm(carb ~ hp + wt + factor(am),
                family="poisson",
                data=mtcars)
# Read the model summary
summary(myModel5)
# Read the model diagnostic
plot(myModel5)
# Visualise the observed / fitted values as a table
tibble(observed = mtcars$carb,
        fitted = myModel5$fitted.values) %>% View()
```

2.3 Logistic Regression

Logistic regression can be used if the dependent variable is binary. This refers to the scenario where the outcome can either be Y or $\neg Y$. The model estimates the probability of outcome $P(Y) \in (0, 1)$ using the logistic function as outlined in equation (2.3).

$$\sigma(X) = \frac{1}{1 + e^{-X}} \quad (2.3)$$

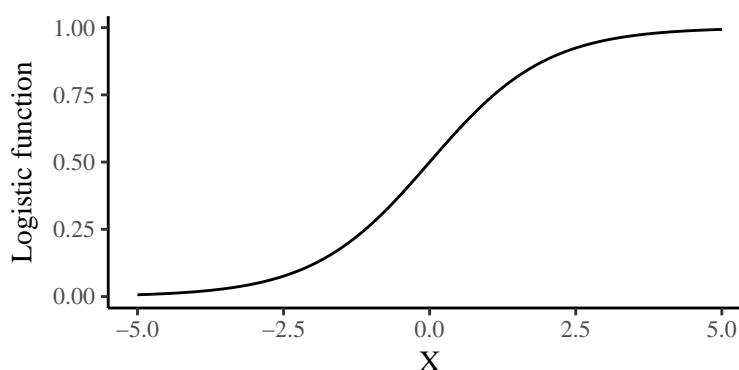


Figure 2.5: Graph showing the range of a logistic function

In a univariate scenario, the logistic function can be expressed as the following equation, where β_0 represents the intercept while β_1 refers to the coefficient of

variable x_1 . The output $\sigma(X)$ indicates the probability of when the label is true, which is always within the range $\sigma(X) \in (0, 1)$. If the model is multivariate in nature, independent variables and their corresponding coefficients can be chained linearly in the X term (2.4).

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_M x_M)}} \quad (2.4)$$

One of the most powerful features of logistic regression is the odds-ratio. It quantifies the effects of each independent variable in a simple way. It is a value indicating the change of likelihood of the event when an independent predictor variable x_1 increase by 1 unit. Odds-ratio is defined in equation (2.5).

$$OR(x_1) = \frac{odds(x_1 + 1)}{odds(x_1)} = \frac{e^{\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \dots + \beta_M x_M}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_M x_M}} = e^{\beta_1} \quad (2.5)$$

Logistic regression can only handle binary problem. If the dependent variable Y has more than two outcomes we can use another algorithm called multinomial logistic regression³. Such problem can also be analysed using artificial neural networks in a much more sophisticated way which we will cover in a later section.

Exercise 6 Building a Logistic Model

In this exercise, we would continue to use the `mtcars` dataset. This time we will build a logistic regression model to predict whether the vehicle has automatic or manual transmission system (using the `am` variable as dependent variable). In the following example, the model has three independent variables `mpg`, `hp` and `disp`. This logistic regression model can be expressed as equation (2.6).

$$P(\text{manual}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{mpg} + \beta_2 x_{hp} + \beta_3 x_{disp})}} \quad (2.6)$$

You may run the following code to build the logistic regression model. You may also calculate the odds-ratio and analyse the effect of each variable. The last part of the code is to calculate the model accuracy.

```
# Build a logistic regression model to predict the dependent variable am
# (1=manual; 0=auto)
myModel6 <- glm(am ~ mpg + hp + disp, family="binomial", data=mtcars)
summary(myModel6)
# Calculate the odds-ratios by taking the exponential of the coefficients
# Can you explain the effects of each of these independent variables?
```

³<http://www.ats.ucla.edu/stat/r/dae/mlogit.htm>

```
exp(myModel6$coefficients)
# You may also calculate the 95% confidence interval of the odds-ratio
exp(cbind(oddsratio=myModel6$coefficients, confint(myModel6)))
# Returns the modelled probability
myProb <- predict(myModel6, mtcars, type="response")
# Turn the probability into a binary class (i.e. TRUE vs FALSE)
# Probability > 0.5 means the vehicle likely to have manual transmission
myPrediction <- myProb > 0.5
# Construct a contingency table to check correct & incorrect predictions
table(myPrediction, observed=(mtcars$am == 1))
# Calculate model accuracy
# (defined as the percentage of correct prediction)
myAccuracy <- sum(myPrediction==(mtcars$am == 1))/nrow(mtcars)
myAccuracy
```

Chapter 3

Tree-based Methods

Tree-based algorithms belong to the supervised learning discipline. For any given set of labelled objects, trees would produce a set of prediction rules. It is based on the concept of region (denoted as \mathcal{R}_i) which refers to a subset of the original object space. Trees would produce a prediction for all objects situating within the same region.

At the heart of tree-based method is a concept called binary recursive partitioning. It is a top-down process which starts from initial object space. At the first recursion, the algorithm would split the master region \mathcal{R}_1 into two at a cut-off point. This produces two corresponding new regions $\mathcal{R}_2 \subseteq \mathcal{R}_1$ and $\mathcal{R}_3 \subseteq \mathcal{R}_1$ with two distinct prediction values. The cutoff point s determines where to slice the master region. All objects within $\{X|X_1 \geq s\}$ belong to \mathcal{R}_2 and those with $\{X|X_1 < s\}$ belong to \mathcal{R}_3 . This process runs recursively until it hits a termination criteria.

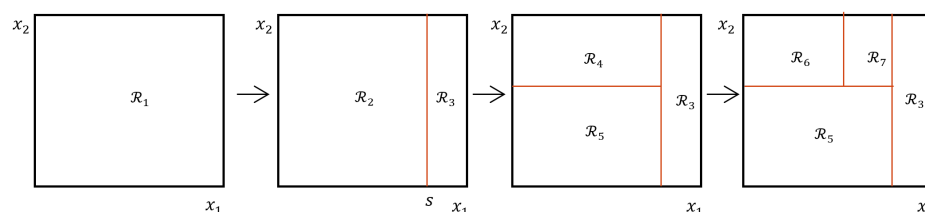


Figure 3.1: Recursive partitioning

3.1 Decision Trees

Decision tree is the simplest form of any tree-based models. It uses standard recursive partitioning to produce prediction regions. Decision tree is a very generic algorithm which fits both regression and classification problem. This means it can predict both continuous real numbers and discrete categories depending on the type of problem. The region prediction for a classification tree is decided by the majority class, while the prediction for a regression tree is defined as the simple average of all members within the region. The tree-splitting structure is called the topology, which can be interpreted graphically in most cases.

On the downside, one of the biggest problems of partitioning is that it tends to produce very deep and complex trees which may risk overfitting the data. Various control parameters can be used to mitigate the overfitting problem. For example, recursion can terminate once all regions are small enough to contain less than 10 objects.

Exercise 7 Growing a Decision Tree

In the R language, there are many packages which implement tree-based algorithm. In this exercise, we will use the `rpart` function in the `rpart` package to build a simple decision trees for a regression problem. The aim of this exercise is to predict car efficiency (`mpg` variable) using the `mtcars` dataset.

```
# Load the rpart package for recursive partitioning
# Load the rpart.plot package for tree visualisation
library(rpart)
library(rpart.plot)
# Build a decision tree to predict mpg using several input variables
myTree <- rpart(formula = mpg ~ wt + hp + factor(carb) + factor(am),
                 data = mtcars,
                 control = rpart.control(minsplit=5))
# Read the tree topology
myTree
# Read the detailed summary of the tree
summary(myTree)
# Visualise the decision tree
rpart.plot(myTree)
```

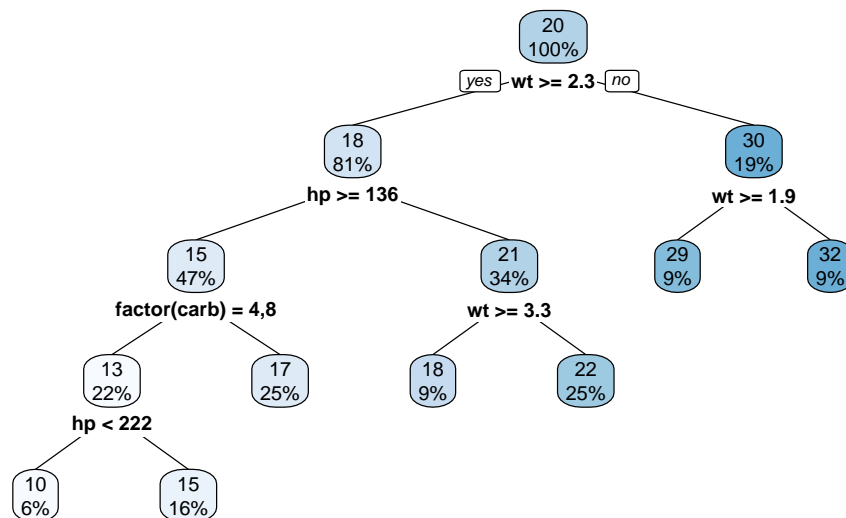


Figure 3.2: Decision tree for a regression problem

Exercise 8 Tree Pruning

It is a common practice to grow a complex tree and then decide how to prune it afterwards. By removing weak branches of the tree, it usually enhance predictive power of the model and eventually makes it more robust.

The command `printcp(myTree)` returns the relative error of the tree at each and every node. It is defined as $1 - R^2$ and therefore always starts with 1 at the top level. As the tree grows, the R^2 value would increase and approach 1 while the corresponding relative error will diminish towards zero. The `cp` value is simply the amount of error reduced when a region is split into two. We can specify a threshold `cp` value so that branches with weak predictive power can be pruned away.

```

# View the cp values
plotcp(myTree)
printcp(myTree)
# Prune the tree at a certain threshold cp value
# You can change the threshold value
myNewTree <- prune(myTree, cp = 0.03)
rpart.plot(myNewTree, fallen.leaves = FALSE)

```

3.2 Random Forest

A random forest is a collection of many small decision trees. Each of the trees are trained using the same dataset, but with randomly selected predictor variables. For a random forest model with P independent variables, only $p < P$ variables are randomly selected for each tree. Such randomness causes variation among the trees. Some trees would have strong prediction power while some others would be weaker.

Once all the trees are grown, the random forest algorithm combines the output of all trees and use the simple average of the region as prediction value if it is regression problem. Alternatively if it is a classification problem, the majority label of the region would become the prediction value.

It is widely recognised that prediction accuracy of random forest is far better than an individual decision tree. However as a trade-off, random forest is often harder to interpret manually as the decision rule becomes more complicated.

Exercise 9 Planting a Random Forest

In this exercise, we would plant a random forest with many individual decision trees. Each tree would randomly select a few variables for consideration. You can use the `randomForest` package to build random forest rapidly.

The importance of each predictor variable is usually indicated by the decrease of node impurity. A powerful predictor would substantially decrease node impurity. For regression, it is measured by residual sum of squares. For classification, the node impurity is measured by the Gini index. You can use the function `importance(myForest)` or `varImpPlot(myForest)` to calculate the importance measurement.

```
library(randomForest)
# The randomForest package does not support factor encoding on-the-fly
# So let us clone the data frame and encode the variable first.
mtcars2 <- with(data = mtcars,
  expr = {carb = factor(carb)
          am = factor(am)
          return(mtcars)})
# Build a random forest with 1000 trees
# Each tree has 2 randomly selected variables
# You can change the parameters
myForest <- randomForest(mpg ~ wt + hp + carb + am,
  ntree = 1000,
  mtry = 2,
  data = mtcars2)
# Plot the error as the forest expands
plot(myForest)
# Plot the distribution of tree size
treesize(myForest) %>% hist()
# Model summary
myForest
# Relative importance of each independent variable
importance(myForest)
varImpPlot(myForest)
```


Chapter 4

Neural Networks

Artificial neural networks (ANN) are mathematical algorithms inspired by the structure of the biological brains. It process incoming information through non-linear mechanism in a neuron and pass on the output to another neuron. When this process repeats many times via multiple layers of neurons, it becomes an artificial neural network. Neural networks having complex structure are usually trained iteratively using backpropagation techniques over long period of time with massive computational power. Nowadays, many modern applications are based on state-of-the-art neural networks, such as video analysis, speech recognition, chatbots and machine translation.

In an ANN, each hidden neuron carries a non-linear activation function $f(x)$. Sigmoid function is a traditional choice of activation function for ANNs(4.1a). It takes the weighted sum of input plus the bias unit and squashes everything into the range $(0, 1)$ with a characteristic sigmoidal 'S' shape. As the sigmoid function is differentiable and easy to compute, it soon becomes a popular choice for ANN activation function. However, it suffers from weak gradient when the input is far away from zero (i.e. the neuron saturates), which makes the ANN learn very slow.

To address the problem of weak gradient, alternative activation functions have been proposed. For instance, the hyperbolic tangent function can be used(4.1b). It shares the same sigmoidal shape but further stretches the output to the range $(-1, 1)$ therefore provides stronger gradient. Yet, the gradient still suffers from saturation when the input is too small or too large.

Different activation functions can provide stronger gradients while maintaining non-linearity. For instance, the softplus function has strong gradient (i.e. unsaturable) for any positive input (4.1c). However, it has been considered computationally costly as it contains logarithm and exponential terms. In light of this, a simplified version call rectified linear unit (ReLU) is usually used instead (4.1d).

The shape of ReLU is very similar to softplus with the exception that it has a threshold at zero. This means only positive input can lead to activation. However, the weighted sum input can change to negative value during training therefore causing the ReLU neuron to cease training. This is called the dying ReLU problem. To avoid this, more advanced activation functions incorporate a very small gradient in the negative range to allow the neuron to recover. The output of common activation functions are visualised in figure 4.1.

Sigmoid activation

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (4.1a)$$

Hyperbolic tangent activation

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (4.1b)$$

Softplus activation

$$f(x) = \ln(1 + e^x) \quad (4.1c)$$

Rectified linear unit (ReLU)

$$f(x) = \max(0, x) \quad (4.1d)$$

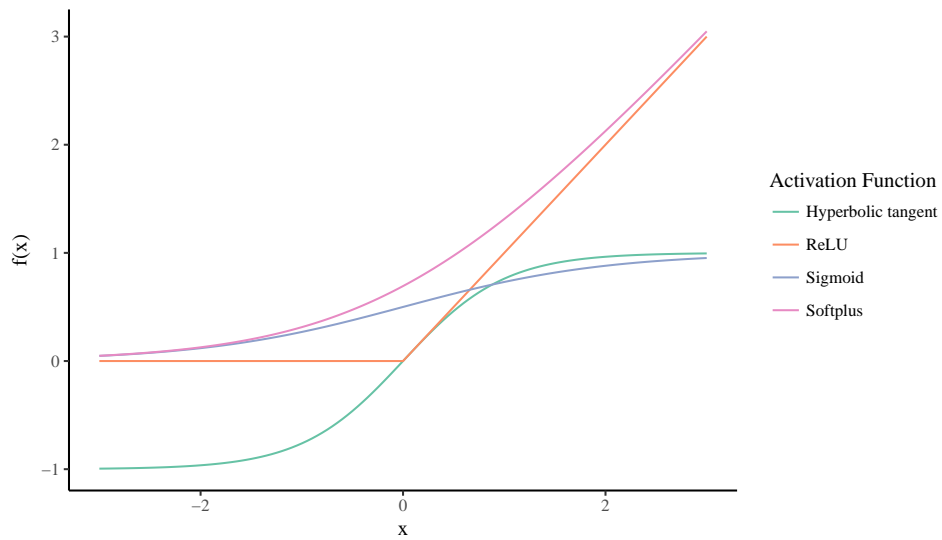


Figure 4.1: Common neural activation functions

The output prediction is made at the network's final layer. Each neuron at this layer combines the hidden neuron's activation through weighted sum and a bias

adjustment unit (4.2a). For regression problems, a linear activation layer is usually used to map the output vector back into unbounded real value range (4.2b). For classification problem, softmax function is used as the final output layer. It maps the hidden layer's activations into the range $(0, 1)$ where the sum of the entire output vector is restricted to 1 in order to represent class probabilities (4.2c).

Weighted sum of hidden vector with bias adjustment

$$y_k = \beta_k + \sum_{h=1}^H w_{h,k} h_h, k = 1, \dots, K \quad (4.2a)$$

Linear output

$$\hat{Y}_k = y_k \quad (4.2b)$$

Softmax output

$$\hat{Y}_k = \frac{e^{y_k}}{\sum_{k'=1}^K e^{y_{k'}}} \quad (4.2c)$$

4.1 Multilayer Perceptron

A multilayer perceptron (MLP) network is the simplest form of all neural network. It consists a configurable number of stacked hidden layers, where every neurons in subsequent layers are fully interconnected.

It is common practice to use zero-centred values for neural network training. In addition, it is vital to check that all variables share the same measurement unit prior to running the model. To achieve this, we would normalise all variables into z -score (4.3a) so that they have the same spread. For any individual value x_i , the z -score can be calculated as the distance from the arithmetic mean \bar{x} divided by standard deviation σ of the variable.

$$z_i = \frac{x_i - \bar{x}}{\sigma} \quad (4.3a)$$

At the training phase, network weights are usually initialised randomly. They are then optimised through backpropagation to achieve gradient descent. The weights improves gradually according to a predefined learning rate until they ultimately converge at the minimum value. One of the drawbacks is that backpropagation does not guarantee reaching global minimum if there are multiple minima across the parameter space. Such problem is usually mitigated by using advanced optimisers with adaptive learning rate.

Exercise 10 Training MLP for Regression Problem

There are many packages which implements neural network in the R language, including `neuralnets`, `nnet`, `RSNNS`, `caret`, `h2o`, `MXNet`, `kerasR`... etc. Generally speaking, all these packages implement the same underlying algorithm and the difference usually lies in syntax, execution speed and hardware compatibility.

In this section, we will continue to use the `mtcars` dataset. The objective of this exercise is to predict the `mpg` value of each car given all other known attributes of it. We would use the `neuralnet` package to create a simple multilayer perceptron (MLP) model.

The following code trains a fully-connected multilayer perceptron with two hidden layers. The code will also visualise the iterative errors, network topology and compute the mean squared error (MSE) for both training and test set.

```
library(dplyr)
library(neuralnet)
# The mtcars dataset has a mixture of numeric and categorical variables
# For numeric variables we need to normalise them
mtcars_numeric <- mtcars %>% select(mpg, disp, hp, drat, wt, qsec)
# foreach numeric variable, we calculate the mean and standard deviation
mtcars_mean <- mtcars_numeric %>% lapply(mean)
mtcars_sd <- mtcars_numeric %>% lapply(sd)
# Convert the numeric variables into z-scores using the mean and sd
mtcars_numeric_normalised <- (mtcars_numeric - mtcars_mean) / mtcars_sd
# Construct a two layers MLP using all numeric variables.
# By default it uses sigmoid active function
myNN1 <- neuralnet(formula = mpg ~ disp + hp + drat + wt + qsec,
  data = mtcars_numeric_normalised,
  hidden = c(4,3),
  linear.output = TRUE,
  lifesign = 'full')
# Visualise the network topology
plot(myNN1)
# Use a helper package for prettier graphics (optional)
library(NeuralNetTools)
plotnet(myNN1)
# Calculate the network prediction
myNNResult1 <- compute(myNN1, mtcars_numeric_normalised %>% select(-mpg))
# The predicted values are in scaled format (z-score)
# Need to convert it back to original scale for comparison
myNNPred1 <- myNNResult1$net.result[,1] *
  mtcars_sd[['mpg']] +
  mtcars_mean[['mpg']]
# Visualise the results on a scatterplot
qqplot(mtcars$mpg, myNNPred1) +
  labs(x='Observed MPG',
  y='Predicted MPG')
# Calculate model error using mean squared error (MSE)
myNNError1 <- mean((myNNPred1 - mtcars$mpg)^2)
```

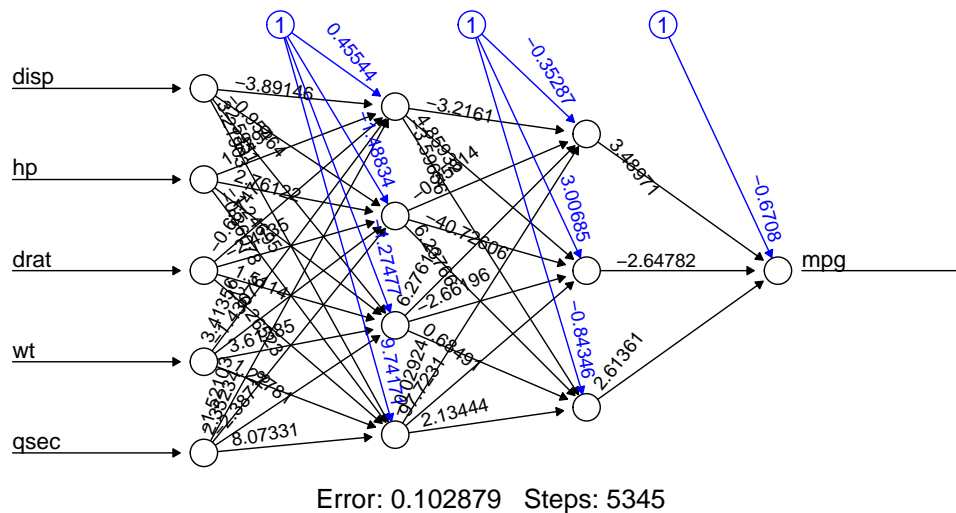


Figure 4.2: MLP model with two hidden layers

In many packages, the default neural activation function is sigmoid function. You can use the following code to use custom activation.

```
# Build the second model
# Keep everything the same but change to softplus activation
myNN2 <- neuralnet(formula = mpg ~ disp + hp + drat + wt + qsec,
  data = mtcars_numeric_normalised,
  hidden = c(4,3),
  linear.output = TRUE,
  act.fct = function(x) { log(1+exp(x)) },
  lifesign = 'full')
# Calculate the network prediction
myNNResult2 <- compute(myNN2, mtcars_numeric_normalised %>% select(-mpg))
# Convert the predicted values back to original scale
myNNPred2 <- myNNResult2$net.result[,1] *
  mtcars_sd[['mpg']] +
  mtcars_mean[['mpg']]
# Calculate model error (MSE)
myNNErr2 <- mean((myNNPred2 - mtcars$mpg)^2)
```

Neural networks can also deal with categorical inputs. They are usually converted into one-hot encoding to feed into the model¹. The following code converts all categorical variables in the `mtcars` dataset into one-hot encoding. The encoded values are then binded to the numeric values and jointly used for network training.

¹For a categorical variable with K unique values, one-hot encoding would produce K new variables. Each new variables would have value $\{1, 0\}$. Please note that this is different from dummy encoding in statistical modelling.

```

# Loads the purrr package to access the imap function
library(purrr)
# Selecting all the categorical variables in the dataset
# Use the imap function to iterate through all columns
# Convert all into one-hot encoding and binds back into a tibble
mtcars_encoded <- mtcars %>% select(cyl,vs,am,gear,carb) %>%
  imap(function(myCol, myName) {
    myUniqueValues <- unique(myCol)
    myTib <- sapply(myUniqueValues,
      function(myValue){ (myValue == myCol) * 1 }) %>% as_tibble
    colnames(myTib) <- paste0(myName, '_', myUniqueValues)
    return(myTib)
  }) %>% bind_cols()
# Combines all numeric and categorical variables
mtcars_all <- bind_cols(mtcars_numeric_normalised, mtcars_encoded)
# View the dataset
View(mtcars_all)
# Train the third model by including encoded categorical variables
myNN3 <- neuralnet(formula = mpg ~
  disp + hp + drat + wt + qsec +
  cyl_6 + cyl_4 + cyl_8 +
  vs_0 + vs_1 +
  am_1 + am_0 +
  gear_4 + gear_3 + gear_5 +
  carb_4 + carb_1 + carb_2 + carb_3 + carb_6 + carb_8,
  data = mtcars_all,
  hidden = c(4,3),
  linear.output = TRUE,
  act.fct = function(x) { log(1+exp(x)) },
  lifesign = 'full')
# Visualise the network topology
plot(myNN3)
# Calculate the network prediction
myNNResult3 <- compute(myNN3, mtcars_all %>% select(-mpg))
# Convert the predicted values back to original scale
myNNPred3 <- myNNResult3$net.result[,1] *
  mtcars_sd[['mpg']] +
  mtcars_mean[['mpg']]
# Calculate model error (MSE)
myNNErr3 <- mean((myNNPred3 - mtcars$mpg)^2)
# Compare the error of the three models
myNNErr1
myNNErr2
myNNErr3

```

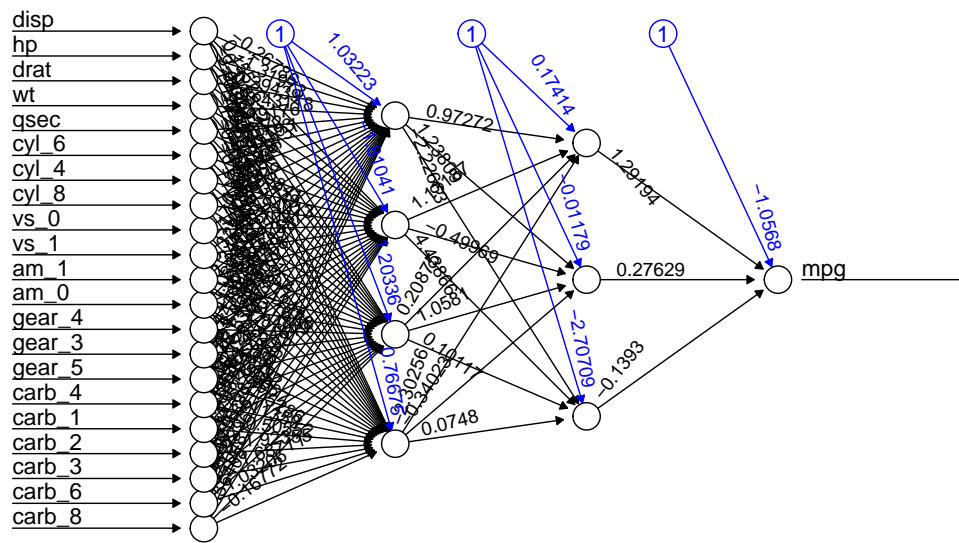


Figure 4.3: MLP model with one-hot encoded categorical variables

Chapter 5

Time Series Analysis

Many datasets have temporal dimension. A time series is a series of observations ordered by chronologically. There are two main types of time series data: 1) regularly-sampled¹, and 2) irregularly-sampled² In this section, we would only focus on time series data sampled at regularly-spaced intervals.

In time series data, temporal properties are often the point-of-interest. That include trend, seasonality, or temporal dependency among different variables. Analysing these properties allows users to gain useful insights. For example, users may extrapolate trend to create forecast going into the future. Alternatively, users may understand the nature of recurring patterns by analysing seasonality. In this section, we will introduce several tools to explore these properties.

5.1 Auto-Correlation Function

Auto-correlation function (ACF) measures the correlation of a single variable along the temporal dimension between x_t and x_{t+h} . In other words, it shows the correlation of the variable over different lag periods.

In the R language, you may use the `Acf(x)` function within the package `forecast` to plot the ACF correlogram. For most time series variables, the correlation is usually strong at lag $h = 1$ and gradually diminishes as lag period increases. Cyclic pattern in the correlogram suggests possible seasonality which you can analyse further.

¹Regularly sampled time series has observations taken at fixed interval. This includes examples like heart rate, network traffic, daily weather... etc.

²Refers to when observations are not recorded at fixed interval, such as incidents of earthquakes.

On the other hand, the partial auto-correlation function (PACF) is similar to the ACF in the sense that it also measures the correlation between different lag periods. The difference is that it controls the correlation across the temporal dimension so that only the contribution of an individual lag can be reflected.

Exercise 11 Loading Datasets

In this exercise, we would use an energy dataset. This dataset contains daily electricity generation and demand data published by a German transmission network called Amprion³. The first column is a date time variable and the rest are demand and generation data, each sampled at 15 minutes granularity. Let us start by loading the dataset from CSV file.

Table 5.1: Description of the Amprion dataset

Variable	Measurement Unit	Description
datetime	%Y-%m-%d %H:%M:%S	Date and time
demand	Megawatt	Demand in control area
pv	Megawatt	Photovoltaic feed-in
wp	Megawatt	Wind power feed-in

```
# Read the Amprion dataset from CSV file
# You might have to modify the path to point to your file location
amprion <- read.csv('amprion.csv',
  colClasses = c('character',
    'numeric',
    'numeric',
    'numeric')) %>% as_tibble()

# View the dataset
amprion

## # A tibble: 140,240 x 4
##       datetime demand    pv    wp
##       <chr>    <dbl> <dbl> <dbl>
## 1 2014-01-01 00:00:00 16231     0 1498
## 2 2014-01-01 00:15:00 16125     0 1449
## 3 2014-01-01 00:30:00 16066     0 1411
## 4 2014-01-01 00:45:00 16137     0 1438
## 5 2014-01-01 01:00:00 16045     0 1256
## 6 2014-01-01 01:15:00 15851     0 1242
## 7 2014-01-01 01:30:00 15729     0 1216
## 8 2014-01-01 01:45:00 15498     0 1312
## 9 2014-01-01 02:00:00 15417     0 1473
## 10 2014-01-01 02:15:00 15471     0 1676
## # ... with 140,230 more rows
```

³Amprion - Demand in Control Area <https://www.amprion.net/Grid-Data/Demand-in-Control-Area/>

One of the main drivers of demand and generation is weather. The weather dataset is published by the Deutscher Wetterdienst⁴. Weather observations are recorded every hour at the Bremen weather station.

Table 5.2: Description of the Breman weather dataset

Variable	Measurement Unit	Description
datetime	%Y-%m-%d %H:%M:%S	Date and time
airtemp	Degree Celsius	Air temperature
sun	Jcm^{-1}	Short-wave global radiation
windspd	$msec^{-1}$	Wind speed
winddir	Bearing	Wind direction
soil10	Degree Celsius	Soil temperature at 10cm depth
soil20	Degree Celsius	Soil temperature at 20cm depth
soil50	Degree Celsius	Soil temperature at 50cm depth
soil100	Degree Celsius	Soil temperature at 100cm depth

```
# Load the Breman weather dataset
bremen <- read.csv('bremen.csv',
                  colClasses = c('character',
                                'numeric',
                                'numeric',
                                'numeric',
                                'factor',
                                'numeric',
                                'numeric',
                                'numeric',
                                'numeric')) %>% as_tibble()

# View the dataset
bremen

## # A tibble: 79,669 x 9
##       datetime airtemp sun windspd winddir soil10 soil20 soil50
##       <chr>    <dbl> <dbl>   <dbl>   <fctr>   <dbl>   <dbl>   <dbl>
## 1 2009-01-01 01:00:00 -2.5    0     3.8    350    -1.8    -1.5    -0.6
## 2 2009-01-01 02:00:00 -2.7    0     3.8    350    -1.7    -1.4    -0.6
## 3 2009-01-01 03:00:00 -2.8    0     1.6     40    -1.7    -1.3    -0.6
## 4 2009-01-01 04:00:00 -2.4    0     1.2    220    -1.6    -1.3    -0.5
## 5 2009-01-01 05:00:00 -2.2    0     1.8    260    -1.5    -1.2    -0.5
## 6 2009-01-01 06:00:00 -1.9    0     2.7    260    -1.5    -1.2    -0.5
## 7 2009-01-01 07:00:00 -1.6    0     2.6    250    -1.4    -1.1    -0.5
## 8 2009-01-01 08:00:00 -1.3    0     3.0    250    -1.3    -1.1    -0.5
## 9 2009-01-01 09:00:00 -0.8    2     2.6    260    -1.1    -1.0    -0.5
##10 2009-01-01 10:00:00  0.0   11     3.0    250    -0.9    -0.9    -0.5
## # ... with 79,659 more rows, and 1 more variables: soil100 <dbl>
```

⁴DWD Climate Data Center https://www.dwd.de/EN/climate_environment/cdc/cdc_node.html

Before moving on to further analysis, we need to normalise the two datasets into the same granularity first. Once this is done, we can then join the two datasets together to form one table containing all variables.

```
# Load the lubridate package to access more datetime functions
library(lubridate)
# Load the dplyr package for data wrangling
library(dplyr)
# Aggregate the amprion dataset from 15 minutes to daily level.
amprion_daily <- amprion %>%
  mutate(date = datetime %>%
    ymd_hms() %>%
    floor_date('day') %>%
    as.Date()) %>%
  group_by(date) %>%
  summarise(total_demand = sum(demand),
            total_pv = sum(pv),
            total_wp = sum(wp))
# Aggregate the bremen dataset from hourly to daily.
bremen_daily <- bremen %>%
  mutate(date = datetime %>%
    ymd_hms() %>%
    floor_date('day') %>%
    as.Date()) %>%
  group_by(date) %>%
  summarise(mean_airtemp = airtemp %>% mean(),
            max_sun = sun %>% sum(),
            mean_windspeed = windspeed %>% mean(),
            mean_soil10 = soil10 %>% mean(),
            mean_soil20 = soil20 %>% mean(),
            mean_soil50 = soil50 %>% mean(),
            mean_soil100 = soil100 %>% mean())
# Join the two daily datasets together into a common table
myTable <- amprion_daily %>%
  left_join(bremen_daily, by = 'date')
# View the aggregated datasets
View(myTable)
# Plots the daily total demand
myTable %>%
  ggplot(aes(x=date, y=total_demand)) +
  geom_line() +
  labs(x = 'Date',
       y = 'Power Demand (MW)')
```

Exercise 12 Analysing Temporal Correlation

Using the code below, you may create an ACF, a PACF and a CCF plot. The latter one refers to the cross-correlation function plot between two time series. In other words, it shows the temporal relationship among two variables across different lag.

```

# Load the forecast package
library(forecast)
# Plots the ACF correlogram only
# There are several ways to create plots.
ggAcf(myTable$total_demand) # More pretty
Acf(myTable$total_demand)   # Standard base R plot
# Plots the PACF correlogram only.
ggPacf(myTable$total_demand)
Pacf(myTable$total_demand)
# Draw a CCF correlogram which find the correlation between two variables.
# You can try swapping variables here.
ggCcf(x = myTable$mean_airtemp,
      y = myTable$total_demand)
Ccf(x = myTable$mean_airtemp,
    y = myTable$total_demand)
# Constructs the several key plots in one go.
ggtstdisplay(myTable$total_demand)
tstdisplay(myTable$total_demand)
# Create a lag plot
gglagplot(myTable$total_demand, lags = 28)
lag.plot(myTable$total_demand, lags = 28)

```

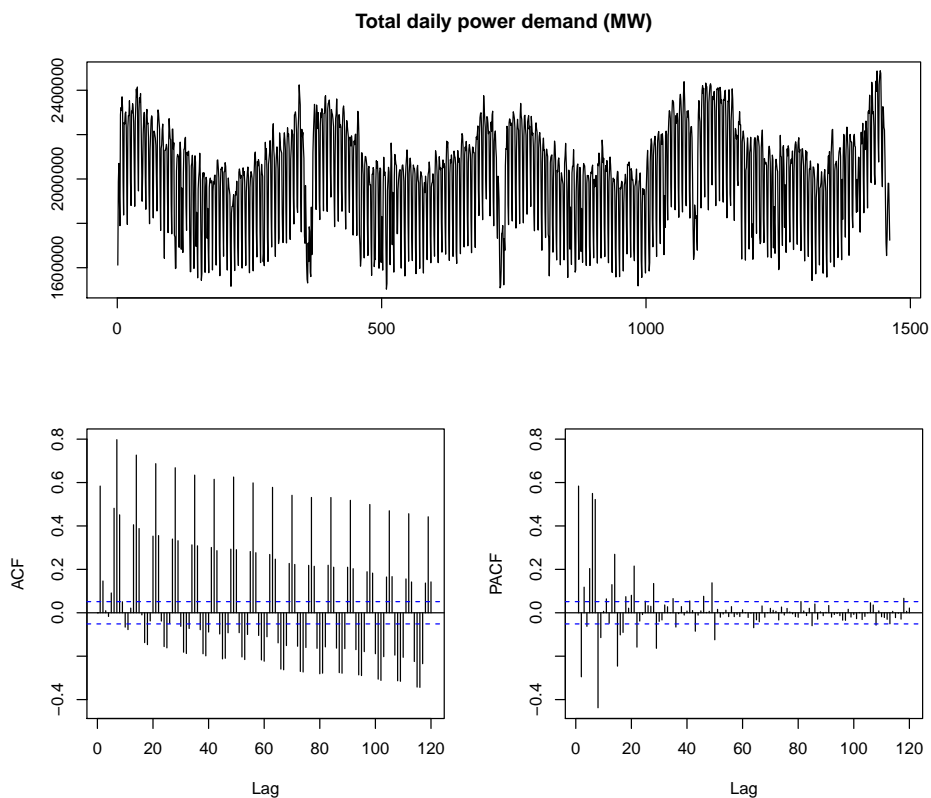


Figure 5.1: ACF and PACF correlograms

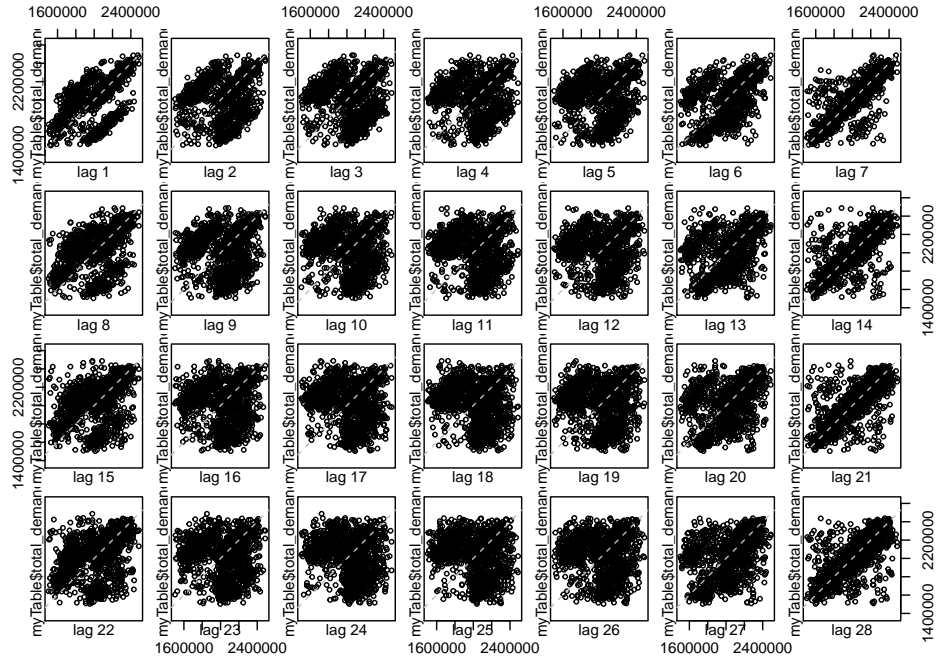


Figure 5.2: Lag plot showing the correlation of various lag periods

5.2 Decomposition

Time series at regular interval can be either additive or multiplicative. Additive time series can be generally described as $X_t = S_t + T_t + \epsilon_t$ where S_t refers to the seasonality at time t while T_t refers to the trend component. The observed data X_t is simply the sum total of trend, seasonal and error components. Alternative, a multiplicative time series is defined as $X_t = S_t \times T_t \times \epsilon_t$. These components can be easily decomposed from the observed values.

Exercise 13 Identifying Trend and Seasonal Components

From this exercise onwards, we would divide the dataset into training and testing set. We would use the training set to build models, the models would be apply to the testing set to assess their performance.

The following code will help us find out the ideal frequency of the seasonal component. It uses spectral analysis to identify frequency with the strongest spectral density. Once the frequency is calculated, we can build a `ts` object using the frequency value. The function `decompose()` would convert the observed time

series into trend component $T_t \in [1, T]$, seasonal component $S_t \in [1, T]$ and random residuals $\epsilon_t \in [1, T]$.

```
# Divide the dataset into training and testing set
TEST_SET_BEGIN <- '2017-01-01'
myTrainingSet <- myTable %>% filter(date < TEST_SET_BEGIN)
myTestingSet <- myTable %>% filter(date >= TEST_SET_BEGIN)
# Automatically search for ideal frequency using training data
# We would expect the frequency to be 7 (weekly pattern)
myFreq <- findfrequency(myTrainingSet$total_demand)
# Check the calculated frequency
myFreq
# Define a seasonal time series object using the frequency value
myTs <- ts(data = myTrainingSet$total_demand,
           frequency = myFreq)
# Decompose the time series into its constituent components
myDecomp <- decompose(myTs,
                      type = 'additive')
# View the decomposed components
autoplot(myDecomp)
plot(myDecomp)
```

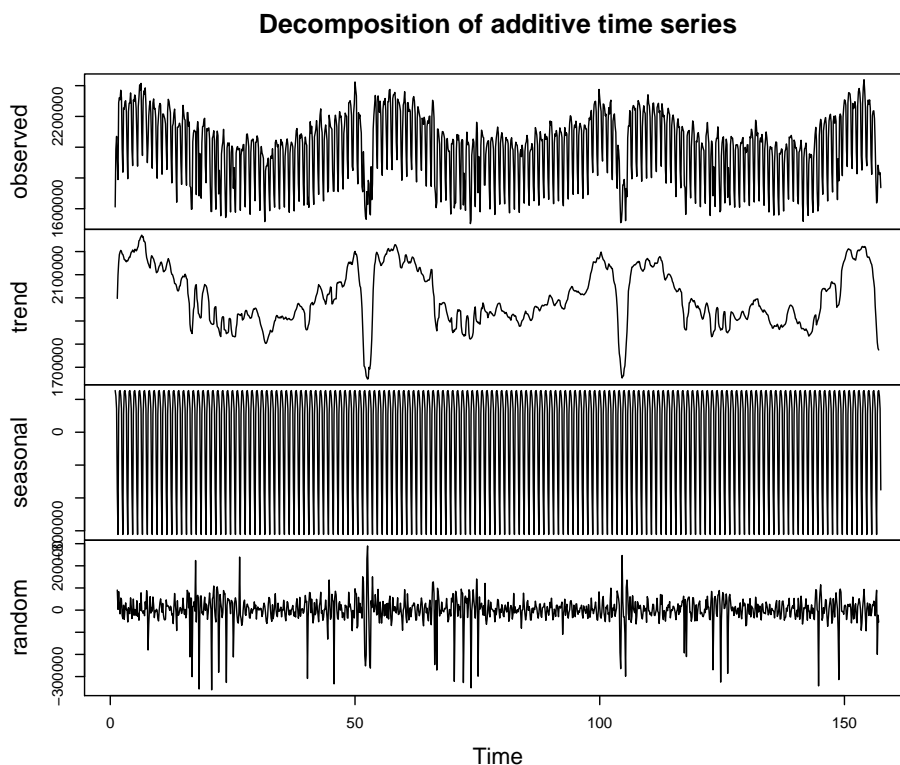


Figure 5.3: Decomposing an additive time series

Exercise 14 Linear Time Series Forecast

Given that you already learnt how to decompose time series data in the previous exercise, we can make further use of the decomposition outputs. In this exercise, we will build a simple forecasting model using the trend and seasonal components as independent variables. The mathematical formula of a linear model with M predictor variables can be expressed as (5.1).

$$X_t = \beta_0 + \beta_{trend}T_t + \beta_{seasonal}S_t + \sum_{m=1}^M (\beta_m x_{mt}) + \epsilon_t \quad (5.1)$$

```
library(forecast)
# Perform linear regression model with decomposed time series components
# You can also add interaction and polynomial terms here
myTsModel1 <- tslm(myTs ~ trend + season +
  mean_airtemp * mean_windspeed +
  poly(max_sun, degree = 2) +
  mean_soil10 + mean_soil20,
  data = myTrainingSet)
# View model summary
summary(myTsModel1)
# Produce forecast using the testing set
myTsForecast1 <- forecast(object = myTsModel1,
  newdata = myTestingSet)
# Visualise the forecast
autoplot(myTsForecast1)
plot(myTsForecast1)
# Calculate the model performance by comparing with the testing set
# Using mean squared error (MSE) here.
myTestError1 <- mean((myTsForecast1$mean - myTestingSet$total_demand)^2)
```

The output of the above code would produce a chart with your forecast visualised as coloured line. The coloured area surrounding the line represents the confidence interval of your prediction.

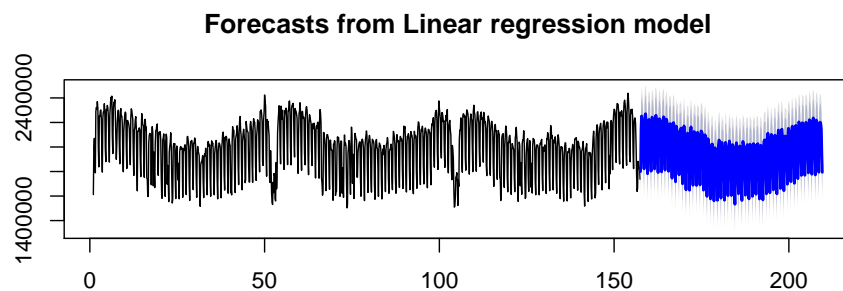


Figure 5.4: Linear time series forecasting with trend and seasonal components

You can run a simple linear regression model using the same set of predictor variables with the `lm()` function to compare with the time series model output.

```
myTsModel2 <- lm(myTs ~ mean_airtemp * mean_windspeed +
  poly(max_sun, degree = 2) +
  mean_soil10 + mean_soil20,
  data = myTrainingSet)
# View model summary
summary(myTsModel2)
# Calculate model prediction using testing set
myTsForecast2 <- predict(object = myTsModel2,
  newdata = myTestingSet)
# Calculate testing MSE
myTestError2 <- mean((myTsForecast2 - myTestingSet$total_demand)^2)
# Compare the testing error (MSE) of these two models
# Which one is a better model?
myTestError1
myTestError2
# Visualise the predictions of the two models
ggplot() +
  geom_line(aes(x=date, y=total_demand), myTrainingSet) +
  geom_line(aes(x=date, y=myTsForecast1$mean, colour='blue'), myTestingSet) +
  geom_line(aes(x=date, y=myTsForecast2, colour='green'), myTestingSet) +
  scale_colour_manual(guide = 'legend', name = 'Model',
    values = c('blue'='blue', 'green'='green'),
    labels = c('Time Series Model', 'Simple Linear')) +
  labs(x='Date',
    y='Total Demand (MW)') +
  theme(legend.position = 'bottom')
```

5.3 ARIMA Model

ARIMA is an acronym for Auto-Regressive Integrative Moving Average model. It is a statistical technique which incorporates lag within the statistical model. It can be described as the combination of three separate parts: autoregression, integration and moving average. ARIMA has three corresponding parameters $p, d,$ and q which is normally expressed as $ARIMA(p, d, q)$ or as separate terms $AR(p)$, $I(d)$ and $MA(q)$.

The $AR(p)$ part suggests that observation X_t is dependent on the linear combination of lagged terms up to p lag periods. A pure $AR(p)$ model is expressed as $X_t = \sum_{i=1}^p (\phi_i X_{t-i})$. The moving average part $MA(q)$ indicates the residual is inherited from up to q lag periods. A pure $MA(q)$ model can be expressed as $X_t = \sum_{i=1}^q (\theta_i \epsilon_{t-i}) + \epsilon_t$. As a result, a simple $ARMA(p, q)$ model can be expressed as the following where ϕ_i and θ_i are model coefficients, while X_{t-i} represent observed data at i^{th} lag step and ϵ_{t-i} refers to the random error at the i^{th}

lag step (5.2).

$$\underbrace{X_t}_{\text{Observation}} = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\sum_{i=1}^p (\phi_i X_{t-i})}_{\text{AR}(p)} + \underbrace{\sum_{i=1}^q (\theta_i \epsilon_{t-i})}_{\text{MA}(q)} + \underbrace{\epsilon_t}_{\text{residual}} \quad (5.2)$$

ARIMA model assumes the time series conforms stationarity⁵. The integrative component $I(d)$ ensures stationarity by taking d number of integrative steps over time. A first order integrative model $I(1)$ is simply the difference between current step and the immediate previous lag step. It is expressed as $X'_t = X_t - X_{t-1}$. Similarly, a second order integrative model $I(2)$ is expressed as $X''_t = X'_t - X'_{t-1} = X_t - 2X_{t-1} + X_{t-2}$.

Time series data with seasonality can be expressed $ARIMA(p, d, q)(P, D, Q)_m$ where the uppercase parameters represent the seasonal component of the model. The m value is a positive non-zero integer indicating the frequency. The estimates $AR(P)$, $I(D)$ and $MA(Q)$ are linearly combined together with the non-seasonal part to create the seasonal ARIMA (SARIMA) model.

Exercise 15 Automated ARIMA

The standard ARIMA implementation accepts six parameters p, d, q, P, D and Q which will produce a seasonal time series model. In many cases, these values are usually not known to the user and all possible values are examined case-by-case to get the best fit.

In the `forecast` package⁶, you may use the function `Arima()` to experiment parameters manually. Alternatively, it is quite common to use automated method to search for good parameters. The method `auto.arima()` tries all parameter values within the given constraints. It can also fit linear regression using predictor variables if the `xreg` attribute is supplied to the function. This is considerably slower than the `Arima()` function due to overhead for parameter search.

⁵A stationary time series has consistent statistical properties across all time, such as equal mean and variance.

⁶The package author has published a detailed book: <https://www.otexts.org/fpp/>

```

library(forecast)
library(dplyr)
# Build an ARIMA model automatically
# Keeping the maximum order (p+d+P+D) small
# Search for seasonal model only
myTsModel3 <- auto.arima(y = myTs,
                        max.order = 5,
                        seasonal = TRUE,
                        xreg = myTrainingSet %>%
                          select(mean_airtemp,
                                mean_windspeed,
                                max_sun,
                                mean_soil10,
                                mean_soil20,
                                mean_soil50,
                                mean_soil100),
                        trace = TRUE)
# View the ARIMA(p,d,q) (P,D,Q) estimates and their coefficients
summary(myTsModel3)
# Run the forecast
# Apply the ARIMA model to testing set
myTsForecast3 <- forecast(myTsModel3,
                          xreg = myTestingSet %>%
                            select(mean_airtemp,
                                    mean_windspeed,
                                    max_sun,
                                    mean_soil10,
                                    mean_soil20,
                                    mean_soil50,
                                    mean_soil100))
# Visualise the forecast
autoplot(myTsForecast3)
plot(myTsForecast3)
# Calculate the MSE error using the testing set
myTestError3 <- mean((myTsForecast3$mean - myTestingSet$total_demand)^2)

```

Exercise 16 Custom ARIMA

After you have calculated the automated ARIMA forecast, you might realise the forecast tends to flat out when forecast horizon increases. This is because the `auto.arima()` function selects the best parameters based on an indicated called AIC. It maximises the log-likelihood of the training data and gives preference to simpler models. We can manually tweak the ARIMA model with custom parameters using the `Arima()` function instead:

```

# Use custom parameters for the ARIMA model
# In this case we can try ARIMA(2,0,0)(1,1,1)
# You can change the parameters here
myTsModel4 <- Arima(y = myTs,
                    xreg = myTrainingSet %>%
                        select(mean_airtemp,
                               mean_windspd,
                               max_sun,
                               mean_soil10,
                               mean_soil20,
                               mean_soil50,
                               mean_soil100),
                    order = c(2,0,0),
                    seasonal = c(1,1,1))

# View the model summary
summary(myTsModel4)
# Apply the ARIMA model to test set
myTsForecast4 <- forecast(myTsModel4,
                           xreg = myTestingSet %>%
                               select(mean_airtemp,
                                      mean_windspd,
                                      max_sun,
                                      mean_soil10,
                                      mean_soil20,
                                      mean_soil50,
                                      mean_soil100))

# Visualise the forecast
autoplot(myTsForecast4)
plot(myTsForecast4)
# Calculate MSE error using the testing set
myTestError4 <- mean((myTsForecast4$mean - myTestingSet$total_demand)^2)

```

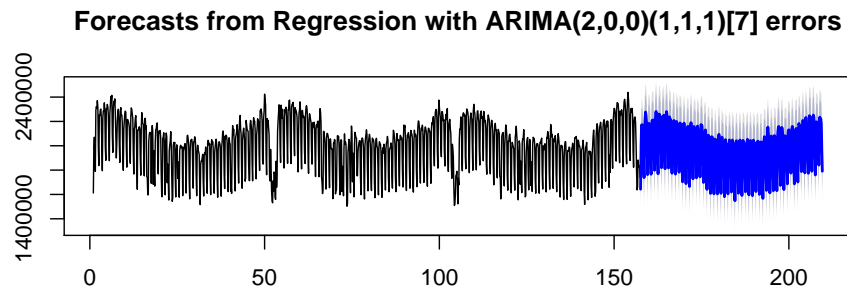


Figure 5.5: Forecast generated from a seasonal ARIMA model.

Exercise 17 Model-based Simulation

With a full $ARIMA(p, d, q)(P, D, Q)_m$ model, users can easily create model-based simulation. The following code will generate 100 simulated runs and plot the average of all runs as point forecast on a chart.

```

# Use one of the trained ARIMA model for simulation
# Wrapping the simulation in a lapply loop
mySimulation <- lapply(1:100, function(i){
  tibble(date = myTestingSet$date,
    run = i,
    value = simulate(object = myTsModel4,
      xreg = myTestingSet %>%
        select(mean_airtemp,
          mean_windspeed,
          max_sun,
          mean_soil10,
          mean_soil20,
          mean_soil50,
          mean_soil100)) %>%
        as.numeric()))))
# Combines all tibbles together to form a large tibble
mySimulationAll <- do.call(rbind, mySimulation)
# Calculate the mean of all simulated runs
myTsForecast5 <- mySimulationAll %>%
  group_by(date) %>%
  summarise(fcast = mean(value))
# Visualise the simulated forecast data
ggplot() +
  geom_line(aes(x=date, y=total_demand), myTrainingSet) +
  geom_line(aes(x=date, y=value, group=run), mySimulationAll, alpha=0.02) +
  stat_summary(aes(x=date, y=value), mySimulationAll,
    fun.y = mean,
    geom = 'line',
    colour = 'blue') +
  labs(x='Date',
    y='Power Demand')
# Calculate the MSE error
myTestError5 <- mean((myTsForecast5$fcast - myTestingSet$total_demand)^2)

```

At last, you can compare the performance of various models:

```

# Linear time series model
myTestError1
# Simple linear regression (not time series model)
myTestError2
# Auto ARIMA model
myTestError3
# ARIMA with custom parameters
myTestError4
# Simulated ARIMA model
myTestError5

```


Chapter 6

Survival Analysis

For problems involving events occurring at irregularly interval, they can be studied through survival analysis. It is commonly used to analyse time-to-event in many research areas, such as medicine, economics, engineering and biology. For example, survival analysis is famously used in clinical research to analyse the effects of different drugs on sustaining patient's life. In this case, the time to death is used an indicator for drug performance. We will go through several techniques in this chapter.

6.1 Kaplan-Meier Estimator

Kaplan-Meier estimator is used to measure how many subjects survives after a certain amount time after treatment started. At time $t \leq T$, the estimator is given by equation (6.1) where $d_{t'}$ represents the number of events and $n_{t'}$ represents the number of subjects at risk.

$$\hat{S}_t = \prod_{t'=1}^t \left(1 - \frac{d_{t'}}{n_{t'}}\right) \quad (6.1)$$

Exercise 18 Fitting a Kaplan-Meier Curve

There are many implementationf for survival analysisin the R language. The most commonly used one is the `survival` package. You can use the `survfit()` function to fit a Kaplan-Meier curve with categorical predictors.

In this exercise, we use the `lung` dataset within the `survival` package which

contains lung cancer patients survival time. You can use the command `?lung` to read the detailed dataset description. To fit a Kaplan-Meier Curve, we need to define the target event (i.e. death, in this example) and the time-to-event. This is done by defining a special object using the `Surv(time, event)` function. The `survfit` function fits a Kaplan-Meier curve against the target event using the supplied variables.

```
# Load the survival package for curve fitting
library(survival)
# Use the survminer package for better graphics
library(survminer)
# Load the lungs dataset into current environment
data(lung)
# Read the dataset description
?lung
# Build an empty model
# We are interested in death cases only (status = 2)
# This model has no predictor variable
mySurvFit1 <- survfit(Surv(time, status==2) ~ 1,
                      data = lung)
# Plot the fitted curve
ggsurvplot(mySurvFit1)
# Use patient's sex as predictor
mySurvFit2 <- survfit(Surv(time, status==2) ~ sex,
                      data = lung)
# Plot the curve with confidence interval and p-value
ggsurvplot(mySurvFit2,
            conf.int = TRUE,
            pval = TRUE)
# The predictor needs to be categorical variable
# Use age as predictor by encoding into age group categories
mySurvFit3 <- survfit(Surv(time, status==2) ~ cut(age, c(40, 50, 60, 70)),
                      data = lung)
ggsurvplot(mySurvFit3, pval = TRUE)
```

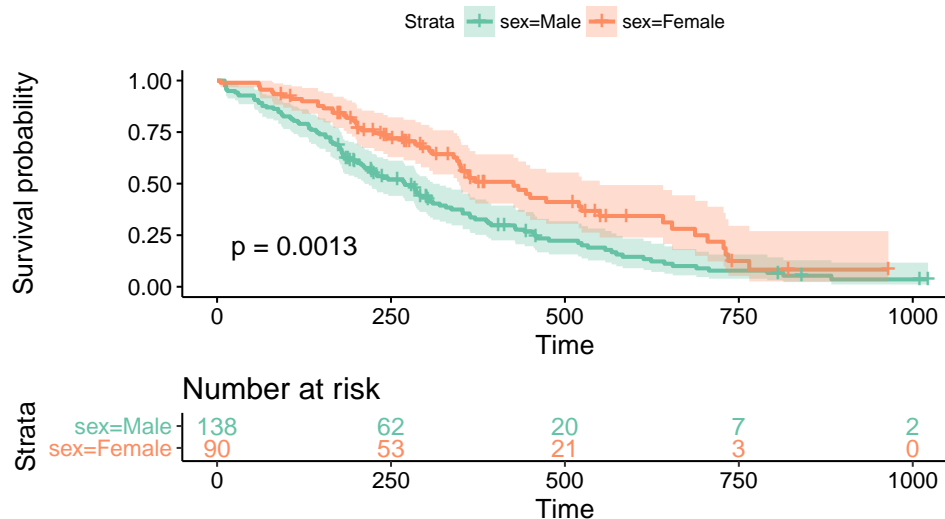


Figure 6.1: Kaplan-Meier curves showing two strata

6.2 Cox Proportional Harzard Model

To investigate the statistical effects of multiple predictor variables on survival probability, a technique named Cox regression can be used. Cox regression can take into account categorical, ordinal as well as numerical range variables. It analyses the simultaneously effects of multiple variables on survival and assumes that the effects of these covariates are time-independent.

The harzard function h_t is give by equation (6.2). The term $h_{0,t}$ in the equation represents the baseline harzard when all covariates are zero. The linear terms $x_1, x_2, x_3, \dots, x_M$ are the predictor variables, while $\beta_1, \beta_2, \beta_3, \dots, \beta_M$ are their corresponding coefficients. For each coefficient β_m , the exponential term e^{β_m} represents the harzard ratio of the covariate variable x_m . If $e^{\beta_m} > 1$, the corresponding covariate is positively correlated with increase in hazard. On the other hand, x_m is negatively correlated with harzard if $e^{\beta_m} < 1$. In the case where $e^{\beta_m} = 1$, the variable x_m has no effects on harzard.

$$h_t = h_{0,t} \times e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_M x_M} \quad (6.2)$$

Cox model assumes variable effects are time-independent. To test whether the assumptions hold, we can analyse the Schoenfeld residuals for each covariate variable. Equation (6.3) defines the Schoenfeld residual $s_{i,k}$ of covariate k of observation i . It is the difference between covariate $x_{i,k}$ and the sum of weighted likelihood of failure of all subjects at risks at time t . If there are observable temporal patterns in the residual plot, it suggests the proportional harzard assumptions may have been violated. In this case, you can consider adding interaction effects with time to mitigate the problem.

$$s_{i,k} = x_{i,k} - \sum_{j \in R(t)} x_{i,m} \hat{p}_j \quad (6.3)$$

Exercise 19 Training a Cox Regression Model

Cox regression model can be trained using the `coxph()` funciton in the `survival` package. In the following example using the `lung` dataset, the cox model analyses the effects of different covariate variables on the time-to-death of a group of cancer patients.

```
# Build a Cox model with predictor variables
myCoxModel1 <- coxph(Surv(time, status==2) ~ factor(sex) + age +
                    ph.ecog + ph.karno +
                    pat.karno +
                    meal.cal + wt.loss, data = lung)

# Read the model summary
summary(myCoxModel1)

## Call:
## coxph(formula = Surv(time, status == 2) ~ factor(sex) + age +
##       ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss, data = lung)
##
##      n= 168, number of events= 121
##      (60 observations deleted due to missingness)
##
##              coef          exp(coef)      se(coef)      z
## factor(sex)2 -0.55085214535    0.57645837470    0.20083299516 -2.74284
## age          0.01064919149    1.01070609595    0.01161113439  0.91715
## ph.ecog      0.73417669164    2.08376570435    0.22327092554  3.28828
## ph.karno     0.02245506374    1.02270907641    0.01123988543  1.99780
## pat.karno    -0.01241655133    0.98766021598    0.00805415700 -1.54163
## meal.cal     0.00003329025    1.00003329081    0.00025946690  0.12830
## wt.loss     -0.01433061217    0.98577158230    0.00777132668 -1.84404
##
##              Pr(>|z|)
## factor(sex)2 0.0060911 **
## age          0.3590623
## ph.ecog      0.0010080 **
## ph.karno     0.0457381 *
## pat.karno    0.1231629
## meal.cal     0.8979096
## wt.loss      0.0651778 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## factor(sex)2 0.5764584  1.7347306 0.3888827 0.8545103
## age          1.0107061  0.9894073 0.9879648 1.0339709
## ph.ecog      2.0837657  0.4799004 1.3452417 3.2277318
## ph.karno     1.0227091  0.9777952 1.0004254 1.0454891
## pat.karno    0.9876602  1.0124940 0.9721916 1.0033750
## meal.cal     1.0000333  0.9999667 0.9995249 1.0005420
## wt.loss      0.9857716  1.0144338 0.9708706 1.0009013
##
## Concordance= 0.651 (se = 0.031 )
## Rsquare= 0.155 (max possible= 0.998 )
## Likelihood ratio test= 28.33 on 7 df,  p=0.0001918421
## Wald test              = 27.58 on 7 df,  p=0.0002615887
## Score (logrank) test = 28.41 on 7 df,  p=0.0001848923
```

The model output above shows the statistical effects of different covariates. For instance, sex^1 is a statistically significant variable for predicting time-to-death of lung cancer patients. This variable has coefficient $\beta_{\text{sex}=2} = -0.551$, which means that having $\text{sex}=2$ would change the patient's hazard by $e^{-0.551} - 1 = -42.4\%$. In other words, $\text{sex}=2$ is beneficial to the patient's wellbeing.

¹It is encoded as male=1 and female=2.

Likewise, `ph.ecog` is also a significant variable. Each unit increase in `ph.ecog` would change the patient's hazard by $e^{0.734} - 1 = 108.4\%$. This implies higher `ph.ecog` score significantly increases patient's risk of death.

Exercise 20 Cox Regression Diagnostics

The code below tests the Cox proportional hazard assumption by calculating the Schoenfeld residuals. If there are observable patterns along the temporal dimension, the model assumptions may have been violated.

```
# Test the proportional hazard assumption of a Cox regression
myCoxZph1 <- cox.zph(myCoxModel1)
# Print the results of the test
myCoxZph1
# Plot the Schoenfeld residuals
ggcoxzph(myCoxZph1)
```

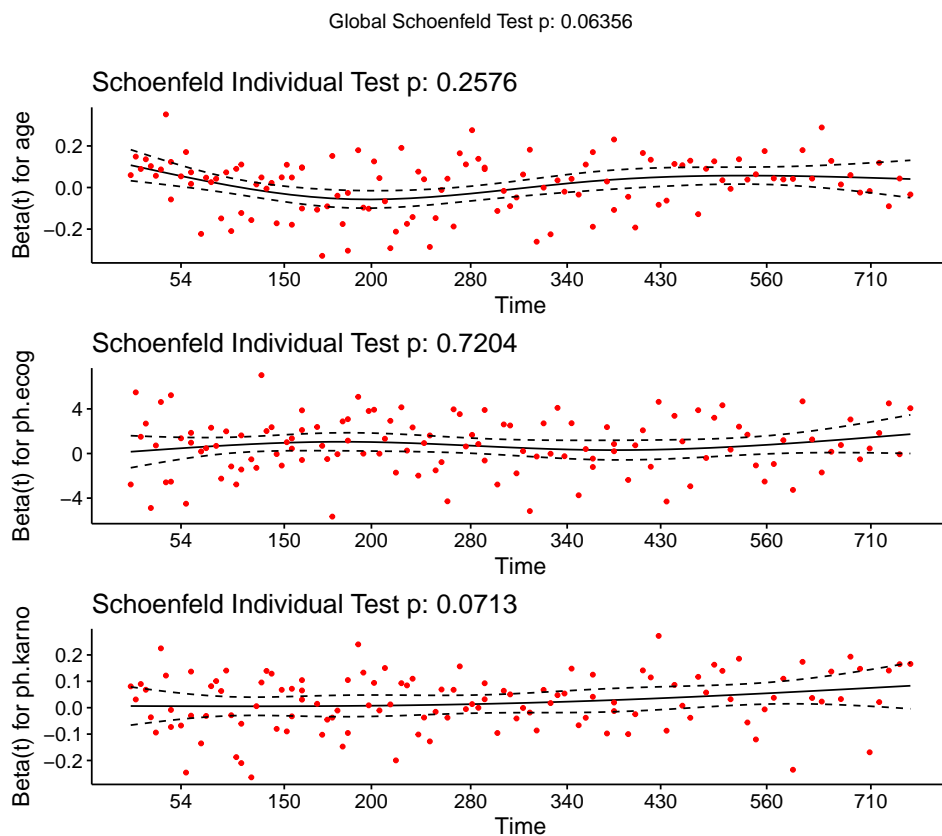


Figure 6.2: Scaled Schoenfeld residuals of a selected set of variables plotted against time

Chapter 7

Unsupervised Learning

Unsupervised learning refers to identifying the underlying structure of an unlabelled dataset. Clustering is one of the most common applications of unsupervised learning, which aims at allocating similar objects into common groups.

In a given set of unlabelled objects, there can be different ways to produce clusters. Figure 7.1 below shows the effects of choosing different number of clusters.

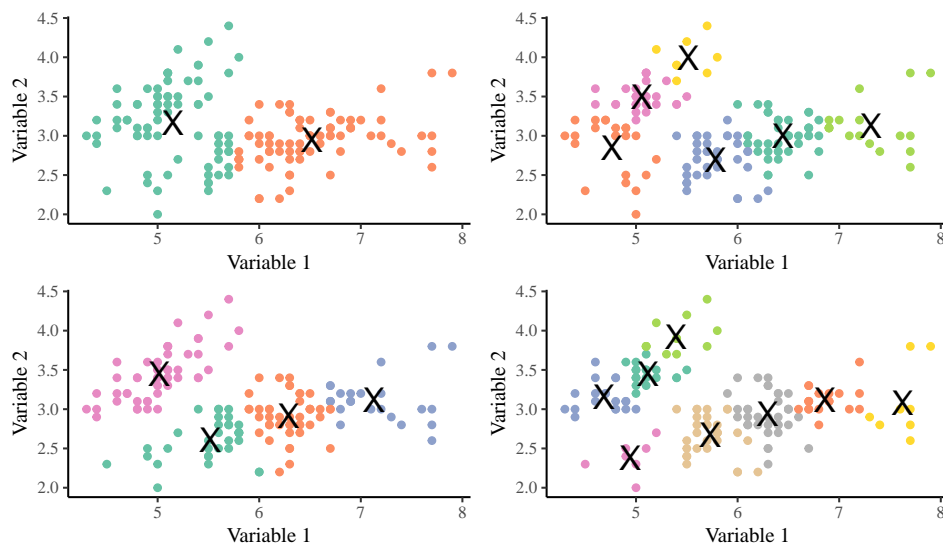


Figure 7.1: Different ways to cluster a set of unlabelled objects

7.1 K -means Clustering

K -means clustering is arguably the most common clustering algorithm due to its intrinsic simplicity. It produces clusters by minimising the Euclidean distance between objects and the centroid of their own cluster. The Euclidean distance between two P -dimensional vectors \vec{x}_j and \vec{x}_k is defined as:

$$d(\vec{x}_j, \vec{x}_k) = \sqrt{\sum_{p=1}^P (x_{j,p} - x_{k,p})^2} \quad (7.1)$$

Algorithm 1: K -means clustering

Input : Set of unlabelled object $X = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_N\}$
Input : Number of clusters K

```

1 Initialise
2  $\{\vec{\mu}_1, \vec{\mu}_2, \vec{\mu}_3, \dots, \vec{\mu}_K\} \leftarrow \text{Randomise}(\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_N\}, K), K < N$ ;
3 while true do
4   for  $k \leftarrow \{1, 2, 3, \dots, K\}$  do
5      $\omega_k \leftarrow \{\}$ ;
6   end
7   for  $n \leftarrow \{1, 2, 3, \dots, N\}$  do
8      $k \leftarrow \underset{k'}{\operatorname{argmin}} d(\vec{\mu}_{k'}, \vec{x}_n), k \in 1, 2, 3, \dots, K$ ;
9      $\omega_k \leftarrow \omega_k \cup \{\vec{x}_n\}$ ;
10  end
11  for  $k \leftarrow \{1, 2, 3, \dots, K\}$  do
12     $\vec{\mu}'_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ ;
13  end
14  if  $\vec{\mu}'_k = \vec{\mu}_k, k = 1, 2, 3, \dots, K$  then
15    break;
16  else
17     $\vec{\mu}_k \leftarrow \vec{\mu}'_k$ ;
18  end
19 end
20 return Cluster centroids  $\{\vec{\mu}_1, \vec{\mu}_2, \vec{\mu}_3, \dots, \vec{\mu}_K\}$ ;
21 return Object cluster assignment  $\{\omega_1, \omega_2, \omega_3, \dots, \omega_K\}$ 

```

Given that the dataset unlabelled, the true number of cluster is not known to us. The K value which represents the number of clusters is usually experimented one-by-one and the best value is determined from the output.

In the R language, the K -means clustering algorithm is implemented very

efficiently. You can use the `kmeans()` function in the `stats` package to perform *K*-means clustering.

Exercise 21 Dimensionality Reduction

In this exercise, we will use the `mtcars` dataset. This dataset contains six numeric variables, in other words, it is a $P = 6$ dimensionality dataset. We can use dimensionality reduction techniques such as principal component analysis (PCA) to visualise them. It converts input variables into principal components (PCs) in the order of maximum variance. The following code would execute PCA and visualise the top two PCs on a scatterplot.

```
# Select the numeric variables from the mtcars dataset
mtcars_numeric <- mtcars %>% select(mpg, disp, hp, drat, wt, qsec)
# Calculate the mean and standard deviation for each variables
mtcars_mean <- mtcars_numeric %>% lapply(mean)
mtcars_sd <- mtcars_numeric %>% lapply(sd)
# Convert the numeric variables into z-scores using the mean and sd
mtcars_numeric_normalised <- (mtcars_numeric - mtcars_mean) / mtcars_sd
# There are six variables in this dataset
# We can use principal component analysis (PCA) to reduce the dimensionality
myPca <- prcomp(mtcars_numeric_normalised)
library(ggfortify)
autoplot(myPca, loadings.label = TRUE)
```

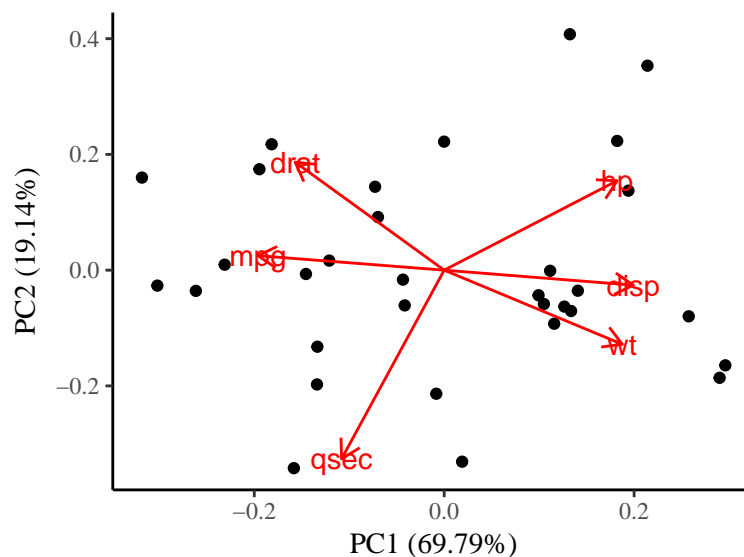


Figure 7.2: Biplot showing the first and second principal components

Exercise 22 *K*-means Clustering

With a $P = 6$ dimensional dataset, we can apply K -means algorithm on it to compute the clusters. In most cases, the number of cluster K is not known in advance. We can experiment different values here.

In practical applications of K -means algorithm, it is very common to normalise the variables using z -scores if they are recorded in different units. Normalisation would ensure that all variables are fairly represented.

```
# We know there are three types of of flowers, so let's start with K=3
# You can try different values
myKClust <- kmeans(mtcars_numeric_normalised, centers = 3)
# Visualise the clusters
ggplot(myPca$x, aes(x = PC1,
                    y = PC2,
                    colour = factor(myKClust$cluster))) +
  geom_point() +
  geom_label(aes(label=mtcars %>% rownames())) +
  stat_ellipse() +
  labs(colour='Cluster')
```

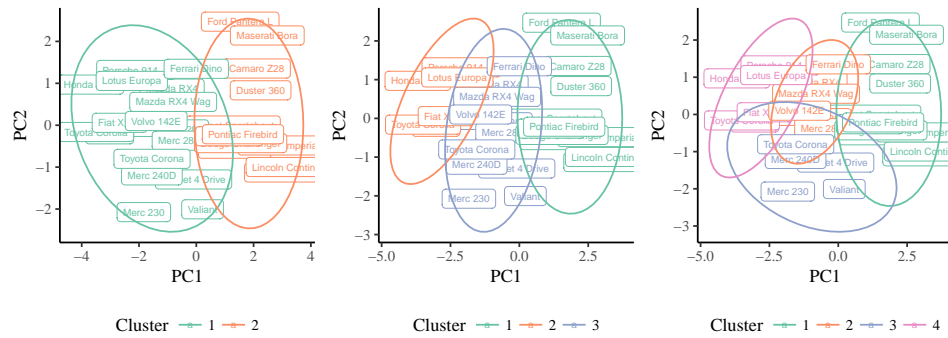


Figure 7.3: Comparing K -means clustering results using different K values

7.2 Hierarchical Clustering

In a given unlabelled dataset containing objects $\vec{x}_i, i = \{1, 2, 3, \dots, N\}$, the maximum number of cluster is N where each cluster contains only 1 member object. In this case, each cluster centroid which is denoted as $\vec{\mu}_i$ contains rich information which perfectly describes its member object as $\vec{\mu}_i = \vec{x}_i$. However, such information would be practically useless. If we start with N clusters, we can merge the closest two clusters into one therefore adding the least amount of error to the system. In this way, we would have obtained $N - 1$ clusters. If we repeat this merging process again and again until there are no more clusters left to merge, you would eventually obtain one single cluster which comprises every member of the object set. This is how agglomerative hierarchical clustering algorithm works.

In hierarchical clustering, the closeness metric between two clusters ω_i and ω_j is denoted as function $D(\omega_i, \omega_j)$. There are several common choices of including 1) single linkage, 2) complete linkage, 3) average linkage, 4) centroid and 5) Ward's method. For single linkage, the distance between two clusters is defined as the closest distance between their member objects (7.2a). In many cases, this tends to produce long chains with similar objects merging sequentially into the same cluster. On the other hand, complete linkage uses the distance between farthest objects between two clusters as the cluster closeness metric (7.2b). This tends to produce clusters with consistent size. The two aforementioned measurements are prone to outlier influence. To mitigate such problem, average linkage can be used. It uses the arithmetic average of all pairwise distance as the cluster closeness measurement (7.2c). Similarly, we can make use of cluster centroid to measure closeness (7.2d). The centroid method is also resilient to outlier influence. Nonetheless, Ward's method compares the change in sum of squares between cluster members and their centroid when they are merged (7.2e).

Single linkage

$$D(\omega_i, \omega_j) = \min_{\vec{x}_i \in \omega_i, \vec{x}_j \in \omega_j} d(\vec{x}_i, \vec{x}_j) \quad (7.2a)$$

Complete linkage

$$D(\omega_i, \omega_j) = \max_{\vec{x}_i \in \omega_i, \vec{x}_j \in \omega_j} d(\vec{x}_i, \vec{x}_j) \quad (7.2b)$$

Average linkage

$$D(\omega_i, \omega_j) = \underbrace{\frac{1}{|\omega_i|} \frac{1}{|\omega_j|} \sum_{\vec{x}_i \in \omega_i} \sum_{\vec{x}_j \in \omega_j} d(\vec{x}_i, \vec{x}_j)}_{\text{Average pairwise distance between } \omega_i \text{ and } \omega_j} \quad (7.2c)$$

Centroid

$$D(\omega_i, \omega_j) = d\left(\underbrace{\left(\frac{1}{|\omega_i|} \sum_{\vec{x}_i \in \omega_i} \vec{x}_i\right)}_{\text{Centroid of } \omega_i}, \underbrace{\left(\frac{1}{|\omega_j|} \sum_{\vec{x}_j \in \omega_j} \vec{x}_j\right)}_{\text{Centroid of } \omega_j}\right) \quad (7.2d)$$

Ward's method

$$\begin{aligned}
 D(\omega_i, \omega_j) = & \underbrace{\sum_{k \in \omega_i \cup \omega_j} \left(\vec{x}_k - \left(\frac{1}{|\omega_i \cup \omega_j|} \sum_{\vec{x}_{k'} \in \omega_i \cup \omega_j} \vec{x}_{k'} \right)^2 \right)}_{\text{Sum of squares of } \omega_i \cup \omega_j} - \\
 & \underbrace{\sum_{i \in \omega_i} \left(\vec{x}_i - \left(\frac{1}{|\omega_i|} \sum_{\vec{x}_{i'} \in \omega_i} \vec{x}_{i'} \right)^2 \right)}_{\text{Sum of squares of } \omega_i} - \\
 & \underbrace{\sum_{j \in \omega_j} \left(\vec{x}_j - \left(\frac{1}{|\omega_j|} \sum_{\vec{x}_{j'} \in \omega_j} \vec{x}_{j'} \right)^2 \right)}_{\text{Sum of squares of } \omega_j}
 \end{aligned} \tag{7.2e}$$

Algorithm 2: Agglomerative hierarchical clustering

Input : Set of unlabelled object $X = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_N\}$
Input : Linkage function $D(\omega_i, \omega_j)$

```

1 for  $n \in \{1, 2, 3, \dots, N\}$  do
2   |  $\omega_n \leftarrow \{\vec{x}_n\}$  ;
3 end
4  $\Omega \leftarrow \{\omega_1, \omega_2, \omega_3, \dots, \omega_N\}$  ;
5 while  $|\Omega| > 1$  do
6   |  $\Omega' = \{\}$  ;
7   | for  $i \in \{1, 2, 3, \dots, |\Omega|\}$  do
8     |  $\Omega'_i \leftarrow D(\omega_i, \omega_j), j = \{1, 2, 3, \dots, |\Omega|\}$  ;
9   | end
10  |  $\{i, j\} \leftarrow \underset{i, j}{\operatorname{argmin}} \Omega'$  ;
11  |  $\omega_{ij} \leftarrow \Omega'_i \cup \Omega'_j$  ;
12  |  $\Omega \leftarrow \Omega' \setminus \Omega'_i \setminus \Omega'_j \cup \omega_{ij}$  ;
13 end

```

The result of hierarchical clustering can be visualised using a tree-like structure called dendrogram. The merging sequence of clusters as well as object closeness can be easily read from the dendrogram. The height of the node at the dendrogram indicates the closeness metric of the two clusters when they are merged. After analysing the dendrogram, users can decide how many clusters to retain. This is usually an objective decision. Once decided, the dendrogram can be cut to obtain the desired number of clusters. Alternatively, we can cut the dendrogram at a certain fixed height to discard trivial clusters at the bottom.

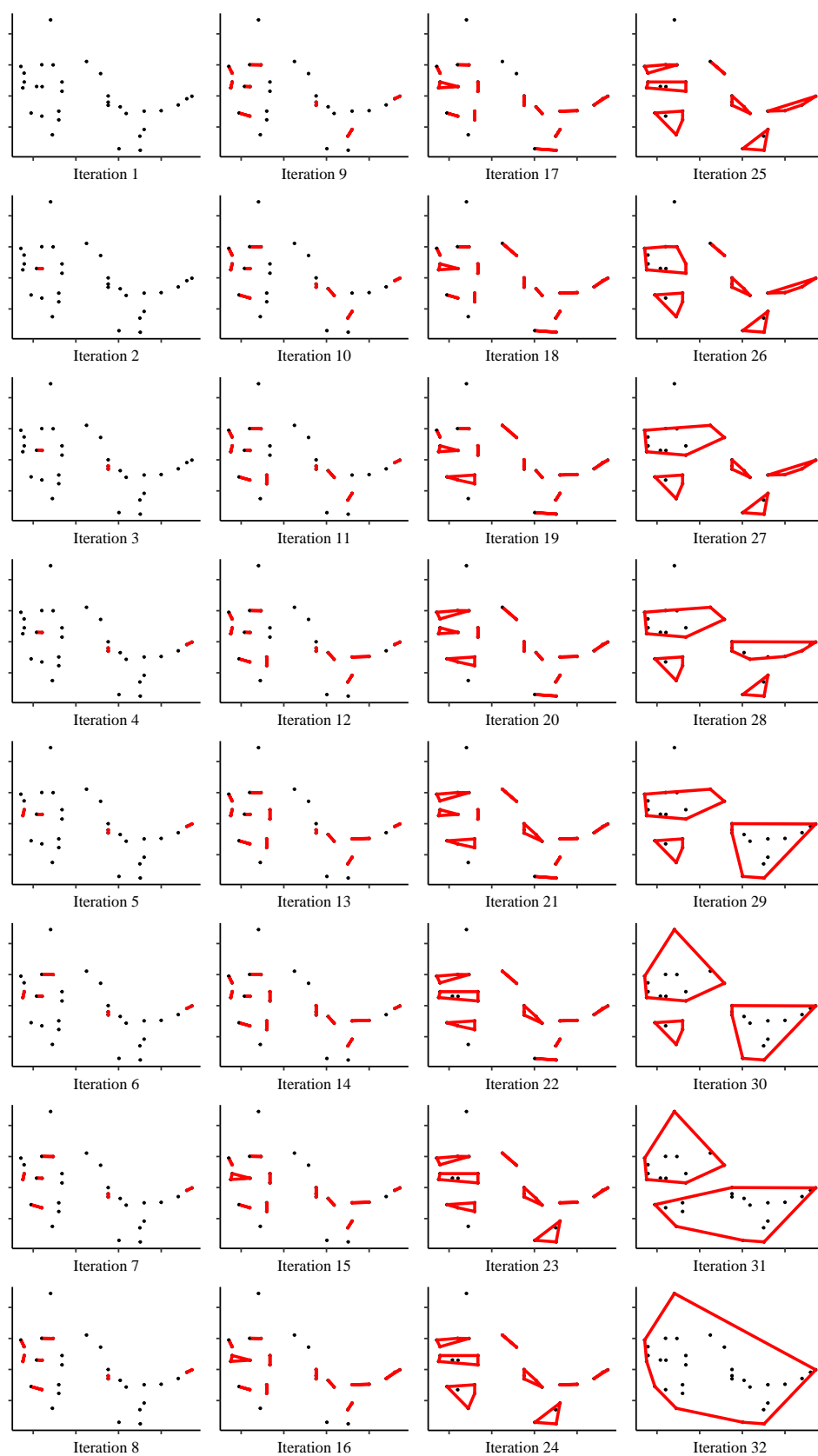


Figure 7.4: Iterative steps of agglomerative hierarchical clustering

Exercise 23 Constructing a Dendrogram

In the R language, hierarchical clustering can be performed using the `hclust()` function which is included in the default `stats` package. The function requires a distance matrix of objects which is normally pre-computed using the `dist()` function. The `hclust()` function uses complete linkage by default if the `method` parameter is not specified. You can change the linkage function and check the difference in output results. The following code snippet produces a simple dendrogram.

```
# Calculate distance matrix
# Using Euclidean distance here but you can change it
myDist <- mtcars_numeric_normalised %>% dist(method = 'euclidean')
# Perform hierarchical clustering using complete linkage
myHClust <- myDist %>% hclust(method = 'complete')
# You can change the closeness measurement
# Read the documentation of the hclust function
?hclust
# Visualise the dendrogram
plot(myHClust)
# You can use gg dendrogram to plot a prettier dendrogram
library(ggdendro)
ggdendrogram(myHClust)
```

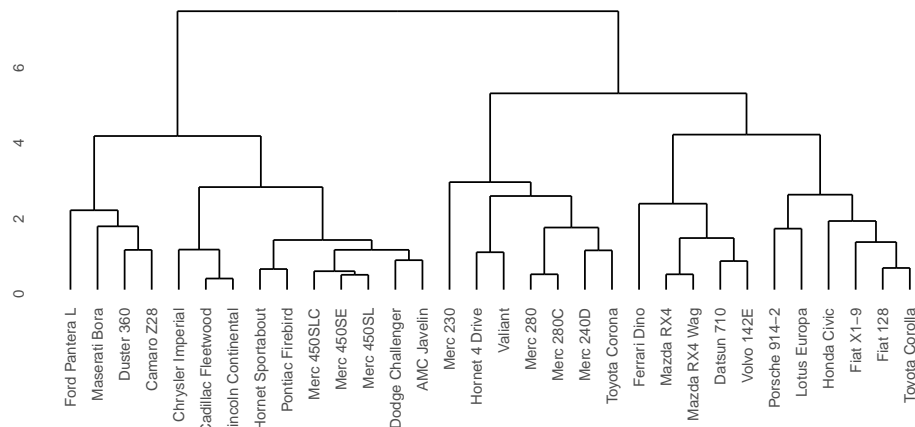


Figure 7.5: Dendrogram illustrating hierarchical clustering using complete linkage

Exercise 24 Cutting a Dendrogram

Dendrogram can be cut to remove smaller clusters at lower height. This can be achieved using the `cutree()` function in R. You can either specify the number of clusters to retain using parameter `k`, this would retain the top `k` clusters of the dendrogram. Alternatively, you can use the `h` parameter to specify at which height the dendrogram should be cut. You can use the code below to cut the dendrogram and visualise the results.

```
# Cut the dendrogram by specifying how many clusters to retain
# You can change the number of clusters here
myCutClusters1 <- cutree(myHClust, k = 5) %>% factor()
# Alternatively, cut the dendrogram at a certain height
myCutClusters2 <- cutree(myHClust, h = 6) %>% factor()
# Use the ape package to plot pretty dendrograms
# The RColorBrewer package generates colour palette
library(ape)
library(RColorBrewer)
# Obtain colour definition
myColours <- brewer.pal(n = 5, name="Set1")
# Convert the hierarchical cluster result into a phylogram object
myPhylo <- myHClust %>% as.phylo()
# Draw some plots
# This is a phylogenetic tree
plot(myPhylo,
     type = 'phylogram',
     tip.color = myColours[myCutClusters1])
# This is a cladogram
plot(myPhylo,
     type = 'cladogram',
     tip.color = myColours[myCutClusters1])
# This is a unrooted phylogenetic tree
plot(myPhylo,
     type = 'unrooted',
     tip.color = myColours[myCutClusters1])
# This is a fan phylogram
plot(myPhylo,
     type = 'fan',
     tip.color = myColours[myCutClusters1])
# This is a radial phylogram
plot(myPhylo,
     type = 'radial',
     tip.color = myColours[myCutClusters1])
```

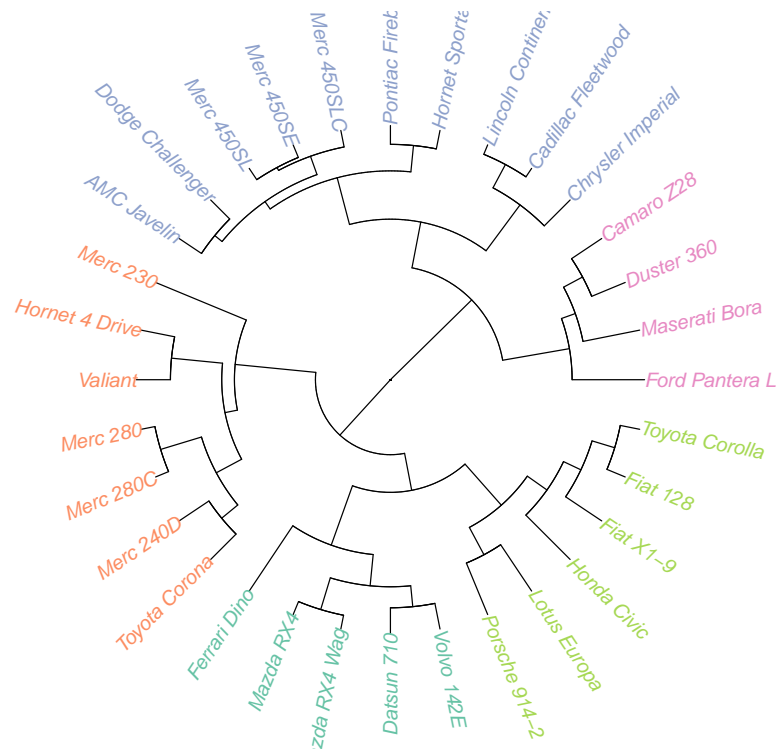


Figure 7.6: Fan phylogram showing hierarchical clusters