People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University Mohamed Kheider of Biskra

Faculty of Exact Sciences and Sciences of Nature and Life

Mathematics Department

Thesis submitted in order to obtain the Diploma :

MASTER in Mathematics

Option: **Statistics**

Submitted by:

ZOUAOUI Nour el Houda

Title :

Parametric Regression Estimation

Members of the Jury :

Pr.	**BRAHIMI Brahim**	UMKB	President
Dr.	**YAHIA Djabrane**	UMKB	Supervisor
Dr.	**BENATIA Fatah**	UMKB	Examiner

Juin 2018

Dedication

This master's thesis is dedicated to :

The greatest hearts in the world, the reason of happiness in my life.

The most wonderful persons in presence, who encourage me, support me, and believe in me.The persons who taught me that there is no such thing called impossible, to :

My parents "my mother and my father".

My brothers.

My sisters

To all my family, the symbol of love and giving.

To my friends.

To all my colleague from 2014 – 2018 *promotion.*

All the people in my life who touch my heart.

Zouaoui Nour el houda

Acknowledgement

First of all, I must thank ALLAH, for giving me the ability, strength, health, and perseverance the necessary to accomplish this work. All praise is due to Allah, the Lord of the Worlds. Peace and blessings be upon our prophet Mohammed.

I would like to express my deep, sincere gratitude to my supervisor DR. YAHIA Djabrane for the confidentiality he has given me by agreeing to frame me. I thank Him deeply for his wise counsel, encouragement, patience, for the continuous support, for being my teacher who taught me with all the splendor, thank you for being a wonderful head of department and for giving us all your experience, motivation and immense knowledge over 2014-2018.

"May Allah reward you for your help."

With great honor, I would like to express my thanks and gratitude to the members of the jury, Pr. BRAHIMI Brahim and DR. BENATIA Fatah, to have examined and evaluated my work.

I would like to give thanks to DR.SAYAH Abdallah for his contribution and assistans through out this humble work. I would also like to give thanks to my dear friend's Sellam Mounira and Mancer Saida for their support and help during the course of this study.

A special thank for all the employees from The Head of the Department of Mathematical to the security agents.

Finally, I must express my profound gratitude to my parents and to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this master's thesis.

This accomplishment would not have been possible without their aid.

Thank you.

Contents

List of Figures

List of Tables

Introduction

In statistics, estimation is an important process of finding the approximate value of some population's parameters from random samples of the population. In this master's thesis, we are interested in the study of parametric estimation in the field of regression; Sir Francis Galton was the first who coined the term "regression". He tried to describe a biological phenomenon, but his work was later extended by Undy Yule, Karl Pearson, Lagendre and Gauss.

Regression analysis is one of the most commonly used statistical methods in practice. The applications of regression analysis can be found in many scientific fields including medicine, biology, agriculture, economics, engineering, sociology, geology, etc. It consists of techniques for modeling the relationship between a dependent variable and one or more independent variables.

In regression, the dependent variable is modeled as a function of independent variables, corresponding regression parameters (coefficients), and a random error term. The parameters of the regression models can be estimated using different method, one of the most commonly techniques is the ordinary last squares (OLS) method.

The main objective of this master's thesis is to study the linear regression, which requires the model to be linear in regression parameters although, we gave a small overview of the nonlinear regression where we discussed some of the popular transformable nonlinear regression models.

Our master's thesis is divided into three chapters. In the first chapter, we tackles the simple linear regression, which aims to model the linear relationship between two variables; one of them is independent variable while the other is dependent. Also we have discussed the very basics characteristics of it.

The second chapter outlines the multiple linear regression that focuses on the linear relationship between one dependent variable and more than one independent variable, we end this chapter with a brief explanation of the nonlinear regression model, which assumes that the relationship between the dependent variable and the independent variables is not linear in regression parameters. In particular, we introduce the transformable models with an explanatory example using the Statistical Package for Social Science (SPSS).

Finally, the third chapter presents a time series regression, which is about the global temperature and we aim to study how the temperature change over years with the use of a polynomial regression model. All the result are required from the software R.

We hope that our audiences will have a beneficial view about regression.

Chapter 1

Simple linear regression

This chapter is devoted the basic idea of linear regression. We started it by giving an introductive example about simple regression and we discussed the estimation of regression parameters containing a single variable.

In addition, we present tests of hypothesis and the confidence interval; moreover, we summarize the analysis of variance by presenting the variance analysis equation.

We explain the determination, correlation coefficient and we end this chapter by defining the most important issue in regression, which is prediction.

1.1 Introduction

One of the oldest topics in the area of mathematical statistics is to find and investigate the relationship between variables. For example, does consumption effect the production? Following the table 1.1 Summarizes a study carried out by Sudanese Company[1]. The data concerns 14 years from 1973 to 1987 about the production and the consumption of a sugar (Sugar was measured in tons "t").

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
x	16.5	17.7	19.0	20.8	22.8	26.3	30.60	34.5	40.3	42.0	47.2	50.3	59.0	66.5
y	382.7	413.2	446.5	466.8	487.8	500.0	520.0	520.0	602.0	613.0	638.0	660.2	700.2	800.1

Table 1.1: Production and consumption data

In order to answer the question: does consumption impact production we may like to know the relationship between the amount of sugar production (y) and the consumption (x). This falls into the field of regression analysis.

Graphically, we can represent this data in a scatter-plot as shown in figure 1.1 clearly from the scatter-plot we can observe that there is a relationship between y and x; in other words, the higher the consumption amount, the higher tends to be the production value. Thus, we have plotted a line that describes this relationship, it indicates the general tendency in which production vary with the consumption performance's level. However, in reality; there are many other factors besides of consumption that will affect the production but not included in the model (like price and agricultural area...) it is usually considered to be of a random nature and it has indicated by ε. We can formulate this relationship as follow:

$$y = f(x) + \varepsilon$$

where f is an affine function from $\mathbb{R}$ to $\mathbb{R}$.

Therefore, the best way to present the effect of a quantitative variable on another quantitative variable is to draw a scatter plot and find the line best fit. Generally, we use the statistical simple linear regression method.

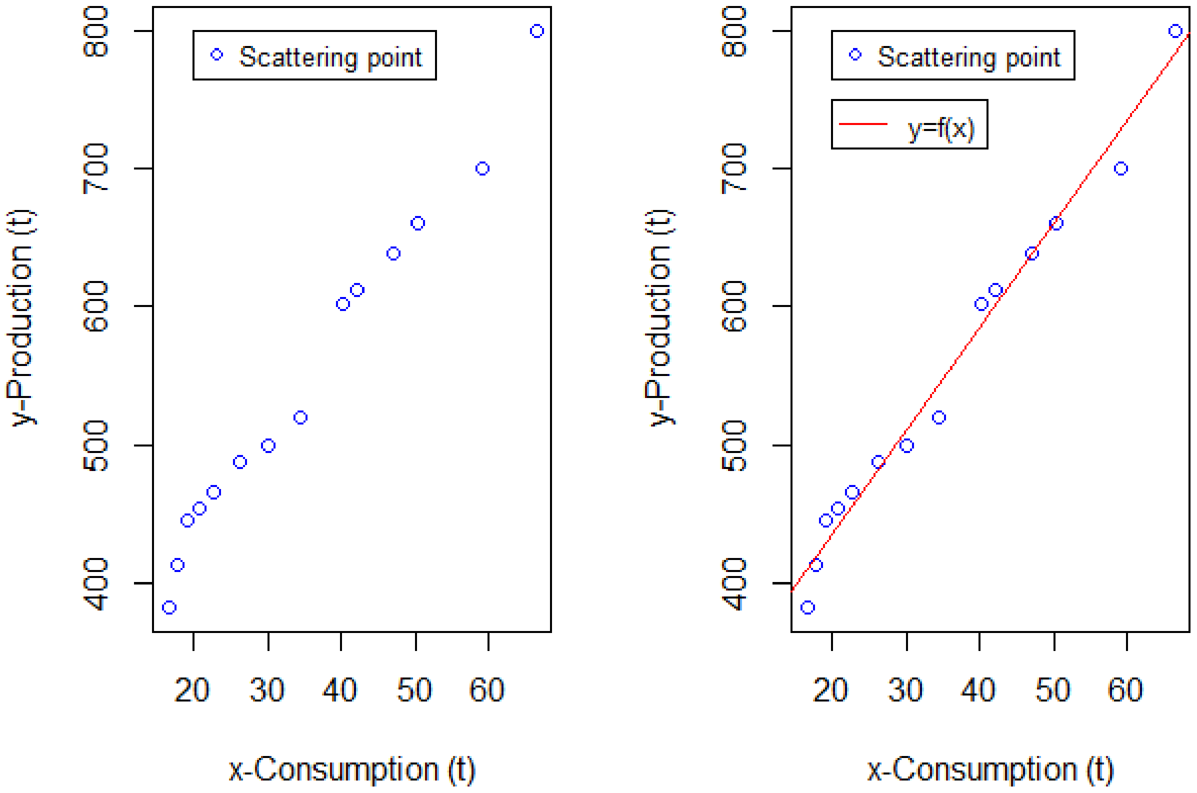

Figure 1.1: Relationship between production and consumption data

1.2 Simple regression model

Definition 1.2.1 *Simple linear regression is used to model the relationship between two quantitative variable. For a set of n observed values* $(x_i, y_i)_{i=\overline{1,n}}$ *of random variable's* (x, y)*, a single response measurement* y *is related to a single predictor* x *for each observation, the model can be expressed as follows:*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , i = \overline{1,n} \tag{1.1}$$

where:

- y_i represents the ith value of the response (dependent) random variable y.

- x_i represents the ith value of the predictor (independent) deterministic variable x.

- β_1 represents the slope of the regression line.

- β_0 represents the intercept of the regression line.

- ε_i is the random error term with mean $E(\varepsilon) = 0$, variance $Var(\varepsilon) = \sigma^2$ and covariance $Cov(\varepsilon_i, \varepsilon_j) = 0$, $\forall i \neq j$.

In addition, those values form a system of linear equations. We can represent it in matrix form as:

$$Y = X\beta + \varepsilon$$

such: $Y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $X := \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\beta := \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ and $\varepsilon := \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

1.3 Parameter estimation

In order to find good estimates for the parameters β_0 and β_1, we employ the Ordinary Least Squares (OLS) method, which gives the line that minimizes the sum squared of the vertical distances from each point to the line.

Remark 1.3.1 *The vertical distance corresponding to the ith observation is:*

$$e_i = y_i - \widehat{y_i}, \; i = \overline{1,n} \tag{1.2}$$

with:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

These vertical distances are called the ordinary least squares residual's. One of the residuals'properties (1.2) is that their sum is equal to zero:

$$\sum_{i=1}^{n} e_i = 0.$$

These residuals can be obtained by rewriting (1.2) as:

$$e_i = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \quad ; i = \overline{1, n} \tag{1.3}$$

The sum of squares of these distances can then be written as:

$$Q(\widehat{\beta_0}, \widehat{\beta_1}) = \sum_{i=1i}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 ; i = \overline{1, n} \tag{1.4}$$

This comes back to determination of minimal optimal of $Q(\widehat{\beta_0}, \widehat{\beta_1})$, that is:

$$\begin{cases} \frac{\partial Q}{\partial \widehat{\beta_0}} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) = 0 \\ \frac{\partial Q}{\partial \widehat{\beta_1}} = -2 \sum_{i=1}^{n} x_i (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) = 0 \end{cases}$$

We find:

$$\begin{cases} \widehat{\beta_0} (\sum_{i=1}^{n}) + \widehat{\beta_1} (\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} y_i \\ \widehat{\beta_0} (\sum_{i=1}^{n} x_i) + \widehat{\beta_1} (\sum_{i=1}^{n} x_i^2) = \sum_{i=1}^{n} y_i x_i \end{cases}$$

Simplifying, we obtain:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{S_{xy}}{S_x^2} \quad \text{and} \quad \widehat{\beta_0} = \overline{y} - \widehat{\beta_1} \overline{x} \tag{1.5}$$

where:

$S_{xy} := \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})$.

$S_x := \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

$\overline{y} := \frac{\sum_{i=1}^{n} y_i}{n}$.

$\overline{x} := \frac{\sum_{i=1}^{n} x_i}{n}$.

Finally, the ordinary least squares regression line is given by:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x \tag{1.6}$$

An important theorem, called the Gauss Markov theorem states:

Theorem 1.3.1 *Under the condition of a regression model (1.1), the least squares estimators β_0, β_1 in (1.5) are unbiased and have a minimum variance among all unbiased linear estimators.[8]*

Remark 1.3.2 *an unbiased estimator of σ^2 is given by :*

$$S^2 := \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2} \tag{1.7}$$

where SSE is the sum of squares of the residuals.

Properties of OLS estimates:

a) $E(\widehat{\beta_0}) = \beta_0$ ($\widehat{\beta_0}$ is unbiased estimator of β_0).

$E(\widehat{\beta_1}) = \beta_1$ ($\widehat{\beta_1}$ is unbiased estimator of β_1).

b) $Var(\widehat{\beta_0}) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] = \frac{\sigma^2}{n}\left[1 + \frac{\overline{x}^2}{S_x^2}\right].$

$Var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2}{nS_x^2}.$

c) $Cov(\widehat{\beta_0}, \widehat{\beta_1}) = Cov(\widehat{\beta_1}, \widehat{\beta_0}) = -\frac{\sigma^2 \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = -\frac{\sigma^2 \overline{x}}{nS_x^2}.$

1.4 Parameter estimation with a normal error distribution

The following parameters β_0, β_1 and σ^2 can be estimated by the method of Maximum Likelihood (ML), when the probability distribution of the error term is determined.

Thus, we assume that the normal error regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i = \overline{1, n} \tag{1.8}$$

where ε_i is independent with normal distribution: $\varepsilon_i \rightsquigarrow N(0, \sigma^2)$ for all $i = \overline{1, n}$ and therefor, $Y_i \rightsquigarrow N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Under the normal errors assumption, the joint density of an observation y_i is:

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{n} f(y_i; \beta_0, \beta_1, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2] \end{aligned} \tag{1.9}$$

Consequently, by using the log-likelihood function we obtain:

$$logL(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

And the first partial derivatives of the log-likelihood function on β_0, β_1 and σ^2 give us:

$$\begin{cases} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = n\sigma^2 \end{cases}$$

Therefore, we get the following estimators:

$$\begin{cases} \widehat{\beta_{1,Ml}} = \frac{S_{xy}}{S_x^2} \\ \widehat{\beta_{0,ML}} = \overline{y} - \widehat{\beta_1} \overline{x} \\ \widehat{\sigma^2}_{ML} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \end{cases}$$

Remark 1.4.1 *We note that the ML estimators of β_0, β_1 are identical to OLS estimators of β_0, β_1, and for $\widehat{\sigma^2}_{ML}$ is biased, therefor, we use the unbiased estimator is given in (1.7).*

1.5 Tests of hypotheses and confidence interval

In this part, we use a formal way of measuring the usefulness of x as a predictor of y, which is a test of hypothesis about the regression parameters.

As stated earlier, we assume that the normal error regression model (1.8) is applicable.

1.5.1 Slope and intercept parameter test

The hypothesis test concerning the slope β_1 and the intercept β_0 is formed respectively:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases} \qquad \& \qquad \begin{cases} H_0: & \beta_0 = 0 \\ H_1: & \beta_0 \neq 0 \end{cases}$$

Since that $\widehat{\beta_1}$ and $\widehat{\beta_0}$ are a linear combination of the observation Y_i, thus, $\widehat{\beta_1}$ and $\widehat{\beta_0}$ will be normally distributed and can be expressed as follow:

$$\widehat{\beta_1} \rightsquigarrow N(\beta_1, \frac{\sigma^2}{nS_x^2}) \qquad \Leftrightarrow \qquad \frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{S_x\sqrt{n}}} \rightsquigarrow N(0,1)$$

$$\widehat{\beta_0} \rightsquigarrow N(\beta_0, \frac{\sigma^2}{n}\left[1 + \frac{\overline{x}^2}{S_x^2}\right]) \qquad \Leftrightarrow \qquad \frac{\widehat{\beta_0} - \beta_0}{\frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{S_x^2}}} \rightsquigarrow N(0,1)$$

Since σ is unknown, we replace it by S given in (1.7), therefore, we obtain:

$$T_{\beta_1} := \frac{\widehat{\beta_1} - \beta_1}{\frac{S}{S_x\sqrt{n}}} \rightsquigarrow t_{n-2} \qquad \& \qquad T_{\beta 0} := \frac{\widehat{\beta_0} - \beta_0}{\frac{S}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{S_x^2}}} \rightsquigarrow t_{n-2} \tag{1.10}$$

Under the null hypothesis, we find:

$$T_{\beta_1} = \frac{\widehat{\beta_1}}{\frac{S}{S_x\sqrt{n}}} \qquad \& \qquad T_{\beta_0} = \frac{\widehat{\beta_0}}{\frac{S}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{S_x^2}}}$$

Accordingly, at the level of significance $\alpha\ \epsilon[0,1]$, H_0 is rejected if:

$$|T_{\beta_1}| \ > \ t_{1-\frac{\alpha}{2}(n-2)} \qquad \& \qquad |T_{\beta_0}| \ > \ t_{1-\frac{\alpha}{2}(n-2)}$$

where: $t_{1-\frac{\alpha}{2}(n-2)}$ is the $(1-\alpha/2)100$ percentile of the student distribution with $(n-2)$ degrees of freedom.

1.5.2 Confidence interval

Since the statistics in (1.10) follow a student distribution respectively, we can make the following probability statement:

$$P(|T_{\beta_1}| < t_{(n-2,1-\frac{\alpha}{2})}) = 1-\alpha$$

$$P(|T_{\beta_0}| < t_{(n-2,1-\frac{\alpha}{2})}) = 1-\alpha$$

Therefore, the $(1-\alpha)$ confidence limits for β_1, β_0 respectively are:

$$\widehat{\beta}_1 \pm t_{1-\frac{\alpha}{2}(n-2)} \frac{S}{S_x\sqrt{n}},$$

$$\widehat{\beta}_0 \pm t_{1-\frac{\alpha}{2}(n-2)} \frac{S}{\sqrt{n}} \sqrt{1+\frac{\overline{x}^2}{S_x^2}}.$$

1.6 The analysis of variance

1.6.1 Variance analysis equation

The variance analysis equation arises from the description of the deviation of the y_i around their mean $\overline{y}$. Thus, we have:

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{:=SST}=\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}_{:=SSE}+\underbrace{\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2}_{:=SSR}$$

such as:

- SST is Total sum of squares.
- SSR is Regression sum of squares.
- SSE is Error sum of squares.

1.6.2 The analysis of variance table for simple regression

The sums of squares are usually laid out in the following analysis of variance table (ANOVA), this table adds a few extra columns like: Degrees of freedom (**Df**)**,** Mean sums of Squares (**MS**) and F-ratios(**F**)**.** The calculations are displayed as follow:

Variation	**SS**	**Df**	**MS**	**F**
Regression	$SSR=\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})^2$	1	$MSR=\frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	$SSE=\sum_{i=1}^{n}(y_i-\widehat{y_i})^2$	$n-2$	$MSE=\frac{SSE}{n-2}$	
Total	$SST=\sum_{i=1}^{n}(y_i-\overline{y})^2$	$n-1$		

Table 1.2: ANOVA table for simple linear regression

where:

MSR is the Regression mean squares.

MSE is the Error mean squares.

The analysis of variance provides us with a useful global test (simultaneous test on $\beta = (\beta_0, \beta_1)$) for regression model. This test is denoted by F which called Fisher Statistic:

$$F = \frac{MSR}{MSE}$$

allows to test:

$$\begin{cases} H_0 : \beta_0 = \beta_1 = 0 \\ H_1 : \exists\ \beta_i \neq 0,\ i = \{0, 1\} \end{cases}$$

Accordingly, H_0 is to be rejected if:

$$F > f_{1-\alpha(1,n-2)}$$

So, test statistic is valide at that level, where $f_{1-\alpha(1,n-2)}$ is the $(1-\alpha)100$ percentile of the Fisher distribution with $(1, n-2)$ degrees of freedom.

1.6.3 Determination and correlation coefficients

- The measure R-squared (R^2) is called the coefficient of determination, it can be interpreted as the proportion of the total variation in Y that is accounted by the predictor variable X, we can express it by:

$$R^2 := \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Since $0 \leq SSE \leq SST$, we note that:

$$0 \leq R^2 \leq 1$$

Remark 1.6.1 *A measure of linear association between Y and X when both Y and X are* *random is the coefficient of correlation r. This measure is the signed square root of R^2:*

$$r = \pm\sqrt{R^2}$$

A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative [8]. Thus, the range of r is:

$$-1 \leq r \leq 1$$

1.7 Prediction

An important application of regression model is predicting new observations y corresponding to a specified level of the predictor variable x.

In other words, prediction situation is when we have a new predictor variable and we want to know the corresponding response, but it has not been observed, yet.

For a new predictor x_k model (1.1) can be written as follows:

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$$

with the following hypotheses:

$$E(\varepsilon_k) = 0,\ Var(\varepsilon_k) = \sigma^2,\ \text{for all } i \neq k,\ Cov(\varepsilon_k, \varepsilon_i) = 0.$$

Thus, the predicted value for y_k is:

$$\widehat{y_k} = \widehat{\beta_0} + \widehat{\beta_1} x_k$$

Now we discuss prediction interval on regression prediction, we have:

$$\begin{aligned}
E(y_k - \widehat{y_k}) &= 0 \\
Var(y_k - \widehat{y_k}) &= Var(\beta_0 + \beta_1 x_k + \varepsilon_k - \widehat{\beta}_0 - \widehat{\beta}_1 x_k) \\
&= \sigma^2(1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2})
\end{aligned}$$

Under the normality assumption of the error term, and substituting σ with S, we find:

$$\frac{y_k - \widehat{y_k}}{S\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \rightsquigarrow t_{n-2}$$

Therefore, prediction interval for $(1 - \alpha)$ prediction limits is:

$$\widehat{y_k} \pm t_{1-\frac{\alpha}{2}(n-2)} S\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Chapter 2

Multiple linear regression

In this chapter, we display the basics of multiple linear regression. Firstly, we introduce the model and its matrix notation, then, we suggest the estimates of parameters by the ordinary least squares. Moreover, we discuss the inference of parameters that contains properties of estimates and confidence interval.

In addition, we present the variance results that hold the coefficient of multiple determination, the overall table of variance and the global test, and we continue by review prediction for multiple linear regression. Finally, we introduce nonlinear regression and show a variety of transformable nonlinear regression models.

2.1 Modeling

Multiple linear regression is an extension (generalization) of simple regression where the data consist of n observations on a dependent variable y and p predictor variables x. The model is presented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \tag{2.1}$$

According to (2.1), each observation can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \qquad i = \overline{1,n} \tag{2.2}$$

where,

- y_i is the ith observation of the response variable y.

- x_{ij} , $j = \overline{1,p}$ is the ith observation of the jth predictor x.

- ε_i is the ith error term.

- β_j , $j = \overline{1,p}$ are the slopes and β_0 is the intercept.

2.1.1 Matrix notation

In matrix terms, the model (2.1) is expressed as:

$$Y = X\beta + \varepsilon \Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

with:

- Y is a vector of responses ($\dim(Y) = (n \times 1)$).

- X is a matrix of constants ($\dim(X) = (n \times (p+1))$).

- β is a vector of parameters ($\dim(\beta) = ((p+1) \times 1)$).

- ε is a random vector of errors with mean vector $E(\varepsilon) = 0_{n\times 1}$ and variance-covariance matrix:

$$Var(\varepsilon) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_n, \;\; I_n := \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

2.2 Estimation method

2.2.1 Ordinary least squares estimates

In order to find the estimator $\widehat{\beta}$ of β, we use OLS method, and for that, we minimize the sum of squares of errors:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + ... + \widehat{\beta}_p x_{ip}))^2$$

So, OLS estimate is given by the formula:

$$\widehat{\beta} = (X^t X)^{-1} X^t Y. \tag{2.3}$$

Denoting Q to the sum of squares errors:

$$\begin{aligned} Q(\widehat{\beta}_0, .., \widehat{\beta}_p) &= \sum_{i=1}^{n} e_i^2 \\ &= e^t e \\ &= (Y - X\widehat{\beta})^t (Y - X\widehat{\beta}) \\ &= Y^t Y - Y^t X \widehat{\beta} - \widehat{\beta}^t X^t Y + \widehat{\beta}^t X^t X \widehat{\beta} \end{aligned}$$

Since $Y^t X \beta$ is a symmetric matrix, therefore:

$$Y^t X \widehat{\beta} = \widehat{\beta}^t X^t Y$$

Consequently:

$$Q(\widehat{\beta}_0, .., \widehat{\beta}_p) = Y^t Y - 2\widehat{\beta}^t X^t Y + \widehat{\beta}^t X^t X \widehat{\beta}$$

So, we calculate the first derivate of Q for β :

$$\frac{\partial Q}{\partial \widehat{\beta}} = -2X^tY + 2X^tX\widehat{\beta} = 0$$

Finally, If the inverse of matrix (X^tX) exist ,we find the OLS estimates (2.3).

Theorem 2.2.1 ***Gauss-Markov***

Among the unbiased estimators of β (linear function of Y), $\widehat{\beta}$ is the one that has a minimum variance with respect to each of its components, Therefore, $\widehat{\beta}$ is the Best Linear Unbiased Estimators (BLUE)[8].

Remark 2.2.1 *-Fitted values are:*

$$\widehat{Y} = X\widehat{\beta} = X(X^tX)^{-1}X^tY \tag{2.4}$$

-Residuals are:

$$e = Y - \widehat{Y} = (I_n - X(X^tX)^{-1}X^t)Y$$

Theorem 2.2.2 *Unbiased estimator of the variance σ^2 in the multiple linear regression is given by [8]:*

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{SSE}{n-(p+1)} \tag{2.5}$$

Remark 2.2.2 *By adding the assumption of normality to the error term ($\varepsilon \rightsquigarrow N_n(0_{n\times1}, \sigma^2 I_n)$) we can use The method of ML which leads to the same estimators at (2.5) and (2.3).the ML function in (1.9) generalizes directly for multiple regression as follows:*

$$L(\widehat{\beta}, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - ..... - \widehat{\beta}_p x_{ip})^2]$$

2.3 Inference about parameter

2.3.1 Properties of the estimates

Theorem 2.3.1 *1) The OLS and ML estimators in β are unbiased :*

$$E(\widehat{\beta}) = \beta$$

2) The variance-covariance matrix of β:

$$Var(\widehat{\beta}) = \sigma^2(X^tX)^{-1}$$

However, this estimator is biased, thus, we use the unbiased in (2.5), therefore:

$$Var(\widehat{\beta}) = S^2(X^tX)^{-1}$$

Proof.

1)

$$\begin{aligned} E(\widehat{\beta}) &= E((X^tX)^{-1}X^tY) \\ &= E((X^tX)^{-1}X^t(X\beta + \varepsilon)) \\ &= \beta + (X^tX)^{-1}X^tE(\varepsilon) \\ &= \beta \end{aligned}$$

2)

$$
\begin{aligned}
Var(\widehat{\beta}) &= E([\widehat{\beta} - E(\widehat{\beta})]^2) \\
&= E([\widehat{\beta} - \beta][\widehat{\beta} - \beta]^t) \\
&= E((X^tX)^{-1}X^t\varepsilon\varepsilon^tX(X^tX)^{-1}) \\
&= \sigma^2(X^tX)^{-1}
\end{aligned}
$$

We find the result by replacing σ^2 into S^2.

■

2.3.2 Test for parameters

For $\varepsilon \rightsquigarrow N_n(0_{n\times 1}, \sigma^2 I_n)$ so $\widehat{\beta} \rightsquigarrow N_n(\beta, S^2(X^tX)^{-1})$, we use the hypothesis test to test β_j:

$$
\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0 \end{cases}
$$

So, under the null hypothesis, test statistic is defined by:

$$
T_{\beta_j} = \frac{\widehat{\beta_j}}{\sqrt{v_j}}
$$

with v_j is jth diagonal term of $S^2(X^tX)$ matrix.

Thus, we reject H_0, if

$$
|T_{\beta_j}| \ > t_{1-\frac{\alpha}{2}(n-p-1)}
$$

Here, $t_{1-\frac{\alpha}{2}(n-p-1)}$ denotes the $100(1-\frac{\alpha}{2})$ percentile of the student distribution with $n-p-1$ degrees of freedom.

2.3.3 Confidence interval

Since, $\varepsilon \rightsquigarrow N_n(0_n, \sigma^2 I_n)$, we have:

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{v_j}} \rightsquigarrow t_{1-\frac{\alpha}{2}(n-p-1)}$$

Hence, $(1-\alpha)$ the confidence limits for β_j are:

$$\widehat{\beta}_j \pm t_{1-\frac{\alpha}{2}(n-p-1)} \sqrt{v_j}$$

2.4 Analyses of variance results

2.4.1 Coefficient of multiple determination

As illustrated in simple regression, in multiple regression we have the decomposition formula:

$$\underbrace{||Y - \overline{Y}||^2}_{:=SST} = \underbrace{||Y - \widehat{Y}||^2}_{:=SSE} + \underbrace{||\widehat{Y} - \overline{Y}||^2}_{SSR}$$

Thus, multiple determination's coefficient is based on this decomposition, and it is defined as follows:

$$R^2 := \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{||Y - \widehat{Y}||^2}{||Y - \overline{Y}||^2}$$

Definition 2.4.1 *The adjusted coefficient of multiple determination R_a^2, is defined by:*

$$R_a^2 := 1 - \frac{(n-1)SSE}{(n-p-1)SST} = 1 - \frac{(n-1)||Y - \widehat{Y}||^2}{(n-p-1)||Y - \overline{Y}||^2}$$

This is supposed to adjust the value of R^2 to account for both the sample size and the number of predictors.

2.4.2 The overall ANOVA table:

The ANOVA table is set up as follows:

Variation	**SS**	**Df**	**MS**	**F**
Regression	$SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$	p	$MSR = \frac{SSR}{p}$	$\frac{MSR}{MSE}$
Error	$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Total	$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$		

Table 2.1: ANOVA table for multiple linear regression

2.4.3 Global test

We denote the global test as follows:

$$\begin{cases} H_0: & \beta_0 = ... = \beta_p = 0 \\ H_1: & \exists\, \beta_j \neq 0 \end{cases}$$

We use the test statistic:

$$F := \frac{MSR}{MSE} = \frac{||\widehat{Y} - \overline{Y}||^2/(p)}{||Y - \widehat{Y}||^2/(n-p-1)}$$

Therefore, the decision rule is:

$$\text{If } F > f_{1-\alpha(p,n-p-1)}$$

we reject the null hypothesis, which means the test is significant, with $f_{1-\alpha(p,n-p-1)}$ denote the $100(1-\alpha)$ percentile of the Fisher distribution with $(p, n-p-1)$ degree of freedom.

Remark 2.4.1 *In multiple regression, the F test does not indicate which parameter β_j is not equal to zero, only that at least one of them is linearly related to the response variable.*

2.5 Prediction and prediction interval

2.5.1 Prediction

In multiple regression, we use the equation (2.4) which can be written as follows:

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + ... + \widehat{\beta}_p x_{ip} \quad ; i = \overline{1,n}$$

Thus, for a new predictor $x_k = (1, x_{k1}, ..., x_{kp})$, the predicted value for y_k is:

$$\widehat{y_k} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{k1} + ... + \widehat{\beta}_p x_{kp}$$

2.5.2 Prediction interval

Using arguments which are similar to the ones in chapter 1, prediction interval in multiple linear regression for a given $x_k = (1, x_{k1}, ..., x_{kp})$ is:

$$\widehat{y_k} \pm t_{(n-p-1,1-\frac{\alpha}{2})} S \sqrt{1 + x_k^t (X^t X)^{-1} x_k}$$

For $(1-\alpha)$ prediction limits.

2.6 Nonlinear regression

Nonlinear regression is a powerful tool for analyzing scientific data, especially if you need to transform data to fit a linear regression. In statistics, nonlinear regression is a form of regression analysis in which observational data are modeled by a function, which is a nonlinear combination of the model parameters and depends on one or more independent variables. The objective of nonlinear regression is to fit a model to the data you are analyzing and estimate the possible parameters from the available data, but it is very complicated to apply.

Therefore, there is a class of function that can be linearized by applying the appropriate transformations. In this case, the estimation of the parameters becomes possible.

Definition 2.6.1 *In general, we can state a nonlinear regression model in the form:*

$$y_i = f(x_i, \beta) + \varepsilon_i$$

An observation y_i is still the response given by the nonlinear response function $f(x_i, \beta)$ and the error term ε_i. The error terms usually are assumed to have expectation zero, constant variance, and to be uncorrelated, just as for linear regression models. Parameter vector in the response function $f(x_i, \beta)$ is denoted by β [8].

2.6.1 Transformable nonlinear regression models

The following table represents the most four common nonlinear regression models that can use in transforming our data in order to achieve a linear relationship between the newly transformed variables and parameters which is the purpose of having a linear function in this form:

$$f(x, \beta) = a + bX$$

Model		Variable Transformation	Transformed parameter
Exponential	$y = \beta_0 e^{\beta_1 x}$	$Y = \ln(y), \quad X = x$	$a = \ln(\beta_0), \quad b = \beta_1$
Logarithmic	$y = \ln(\beta_0 x^{\beta_1})$	$Y = y, \quad X = \ln(x)$	$a = \ln(\beta_0), \quad b = \beta_1$
Power	$y = \beta_0 x^{\beta_1}$	$Y = \log(y), \quad X = \log(x)$	$a = \log(\beta_0), \quad b = \beta_1$
Reciprocal	$y = \frac{1}{\beta_0 + \beta_1 x}$	$Y = \frac{1}{y}, \quad X = x$	$a = \beta_0, \quad b = \beta_1$

Table 2.2: The common nonlinear model transformation

Remark 2.6.1 *In this case, we can easily obtain an estimate of the parameters with the Ordinary Least Squares (OLS) method.*

Example 2.6.1 *A mail order company seeking for the relationship between height and volume of shipping boxes that have a square bases and varying height from* 1 *to* 5 *feet (*1 *ft* = 1 *m) [7]. The data are shown in the following table:*

Height (ft)	1	1.5	2	2.5	3	3.5	4	4.5	5
Volume (ft^3)	2	7	16	31	54	86	128	182	250

Table 2.3: Height and volume data of boxes

So as to solve this problem, firstly, we need to observe the scatter plot of this data as it's shown in figure (2.1).

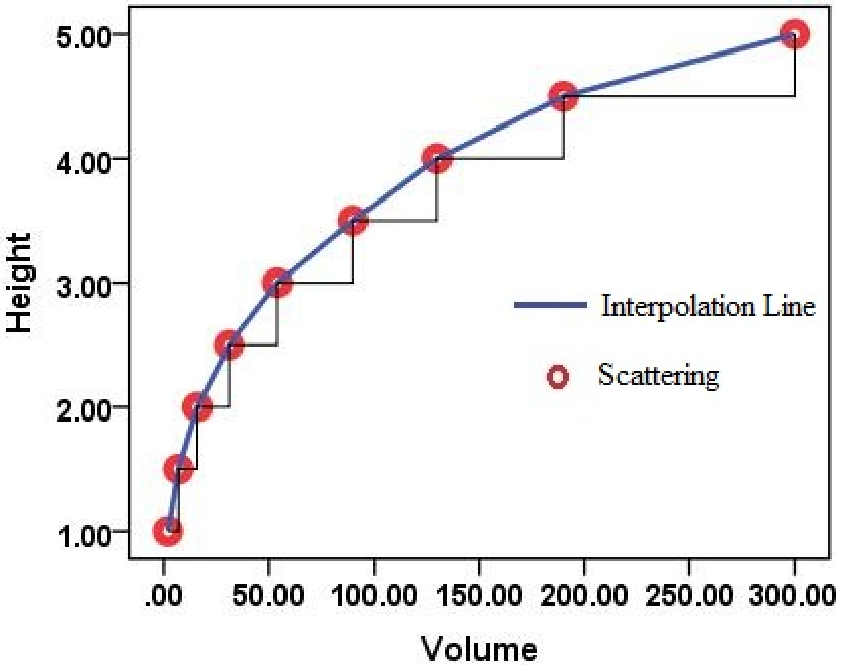

Figure 2.1: Scatter plot of mail order company data & interpolation line.

Graphically, we notice that the scattered and trend curve are clearly nonlinear. Therefore, we need to determine which regression model is most appropriate for data. Manually, we can use the transformation in the table 2.2 to estimate the parameters and also to calculate the R-squared value for each model, then, choose the model with the closest value to one.

In this example, we solve the relationship between height and volume using the SPSS Statistics Software. Particularly, we use the curve estimation command.

SPSS output:

Equation	Model summary					Parameter Estimation	
	R square	F	df1	df2	Sig.	Constant	b1
Linear	0.851	39.891	1	7	0.000	1.853	0.013
Logarithmic	0.959	162.678	1	7	0.000	−0.010	0.817
Inverse	0.514	7.395	1	7	0.030	3.530	−6.091
Power	0.999	7523.898	1	7	0.000	0.804	0.327
Exponential	0.681	14.971	1	7	0.006	1.792	0.004

Table 2.4: Model summary and parameter estimates

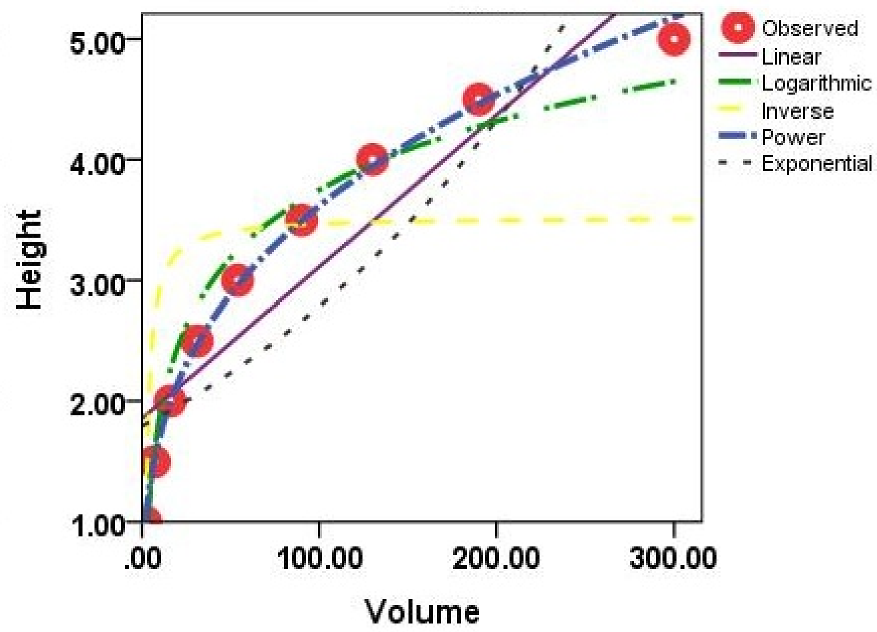

Figure 2.2: The curve fitting for five models

Explanation of the output:

Model summary and parameter estimates table show the result of curve estimation for Linear, Logarithmic, Inverse, Power and Exponential models, it gives the estimated values of regression coefficient (the β coefficient), in addition, the F test values show whether the model is a good fit or not with the p-value and the R-square (R^2 Coefficient of Determination).

The F, df1, df2, and Sig. columns summarize the results of the F test of model fit. The significance value of the F statistic is less than 0.05 *for all models, which means that the variation explained by each model is not due to chance. We note that R square statistic is the important value, which is a measure of association strength between the observed and model-predicted values of the dependent variable (in other words, it is a statistical measure of how close data are to fitted regression line). Thus, the R square for the power model is larger than other values (it is very close to one).*

Finally, analyzing the curve fit chart gives you a quick visual assessment of the fit of each model to the observed values. From this plot, it appears that the Power and Logarithmic model better follow the shape of data. In particular, power model seems the best explanation of the patterns observed in the data with R^2 equal to 0.999.

Therefore, data fits into power curve, the parameter can be obtained from the SPSS output table. In this case, we have $\beta_0 = 0.804$ and $\beta_1 = 0.327$. Thus, regression line of the Power model is:

$$y = 0.804x^{0.327}.$$

Chapter 3

Application

In the previous two chapters, we studied some basic properties of regression analysis. In this chapter, we shed the light on a popular regression model as an example for application.

3.1 Introduction

Time series regression is a statistical method for predicting a future response based on the response history that deals with time series data, which means that data is in a series of particular time periods or intervals.

In this part, we discuss linear regression in the time series context to estimate the parameters of the model. Therefore, to start a time series regression, we should build the model in the sense of regression by assuming the output (dependent) time series, say, y, which being influenced by a collection of possible inputs (independent) series, say, x_t.

We express this relation through the linear regression model:

$$y = f(x_t) + \varepsilon,\ t = \overline{1, n} \tag{3.1}$$

where $f(x_t)$ is the general trend of the model; and the independent error terms ε follow a normal distribution with mean 0 and variance equalto σ^2.

3.2 Representation of the data

Data that will be studied and analyzed is the global temperature anomaly data which come from the Global Historical Climatology Network-Monthly (GHCN-M) data set and International Comprehensive Ocean-Atmosphere Data Set (ICOADS), which have data from 1880 to 2016, that means we have $n = 137$ observations.[9]. For more information on this data, please visit " www.ncdc.noaa.gov/cag/global/data-info ". Data show how variable y which is Global Temperature is changing over time denoted by t. To estimate a time series regression model, a trend must be estimated. We begin by creating a Scatter plot of the time series. Scatter plot can help visualize any relationship between the explanatory variable time (also called regression variable, or predictor) and the response variable Global temperature, it shows how a variable does change over time; it can be used to inspect characteristics of data, in particular, to detect whether a trend exists.

Based on this data, our objective is to build a regression model as shown in (3.1), in which we can set the model up for each observation as follow:

$$y_i = f(t_i) + \varepsilon_i,\ i = \overline{1, n} \tag{3.2}$$

where, the independent variable is time; noted by t and y is the global temperature which is the dependent variable. The following table provides a brief summary of data.

Variable	**Min**	**1st quantile**	**Median**	**Mean**	**3rd quantile**	**Max**
t	1880	1914	1948	1948	1982	2016
y	-0.3006	-0.1774	0.0034	0.0663	0.2273	0.9363

Table 3.1: Summary of the global temperature data

Thus, we plot the yearly global tampreteur using the R code:

>plot(t,y,col='navy',xlab = 'years',ylab = 'Temperature (°C)' ,main='Global temperature anomalies',lwd=2)

The plot looks like this:

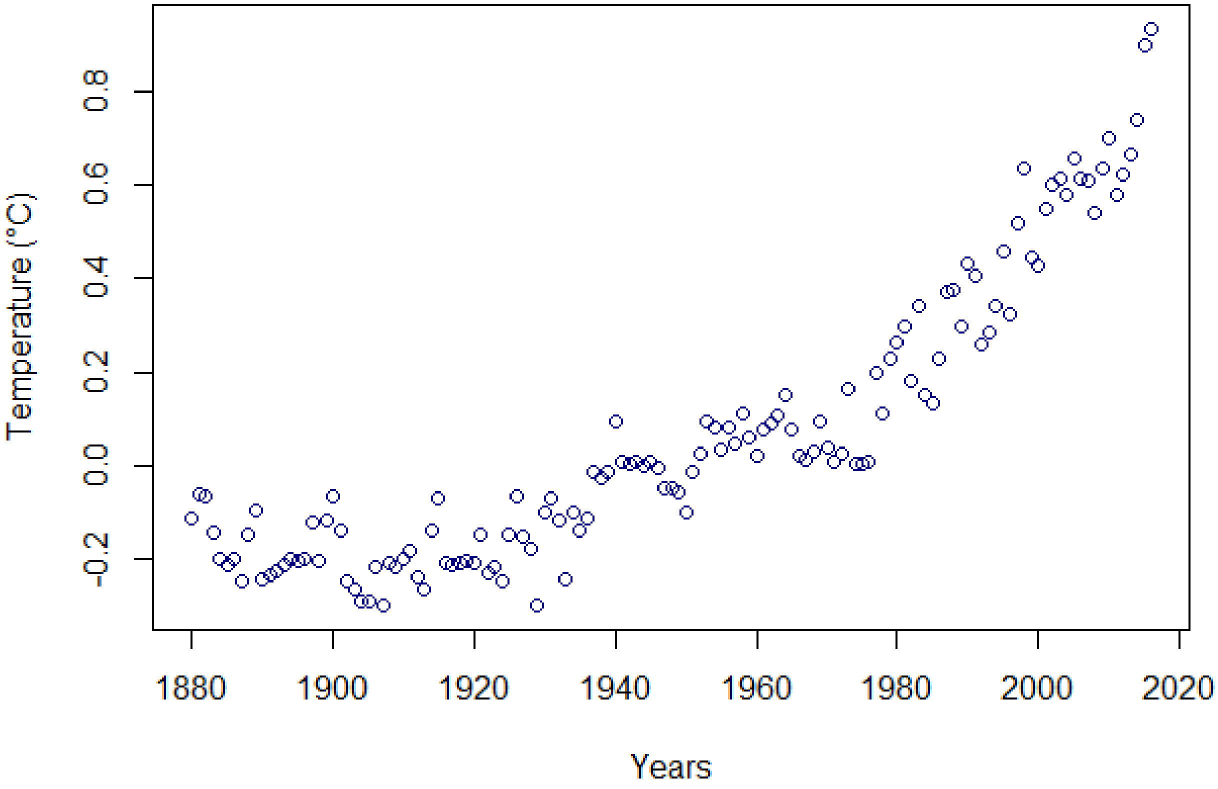

Figure 3.1: Scatter plot of global temperature over 1880-2016

The plot suggests that there is an obvious trend in data; however, the trend does not appear to be quite linear. It appears as if the relationship is slightly curved.

One way of modeling the curvature in these data is by formulating a "polynomial model" with one quantitative predictor, which is the time, it is provided in section (3.3).

3.3 Fitting the model

In this section, we examine the yearly changing of the global temperature and we restrict ourselves to polynomial regression, which is limited from first to fourth degree.

Definition 3.3.1 *Polynomial regression is a form of regression analysis in which we can model the expected value of* y *as an* d*th degree polynomial, yielding the general polynomial regression model,*

$$y = \beta_0 + \beta_1 t + \beta_2 t^2 + ... + \beta_d t^d + \varepsilon$$

with d *is the degree of the polynomial. For lower degrees, the relationship has a specific name (i.e.,* $d = 2$ *is called quadratic,* $d = 3$ *is called cubic,* $d = 4$ *is called quartic, and so on), polynomial regression is still considered linear regression since it is linear in the regression coefficients,* $\beta_0, ..., \beta_d$*, although, polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression [6].*

The matricial form is:

$$Y = T\beta + \varepsilon$$

where

$$Y = X\beta + \varepsilon \Leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \cdots & t_1^d \\ 1 & t_2 & \cdots & t_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \cdots & t_n^d \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Yet, we can define the model that will be estimated by considering the functions in (3.2) of the form:

$$\begin{aligned} f(t_i) &= f(t_i;\beta_0,\beta_1,.....,\beta_d) \\ &= \beta_0+\beta_1 t_i+\beta_2 {t_i}^2+......+\beta_d t_i^d \end{aligned}$$

Thus, the model of temperature data is:

$$y_i=\beta_0+\beta_1 t_i+......+\beta_d t_i^d+\varepsilon_i,\ d=\overline{1,4} \text{ and } i=\overline{1,n}$$

In order to estimate the equation above, we only need the response variable (y) and the predictor variable (t), however, building a polynomial regression model requires estimating model's parameters. Hense, we compare the four possible models and select the best one.

3.3.1 Model building

In order to, formally test whether a linear, quadratic, cubic or quartic trend occurs, we can fit the four models to our data by using the function "lm" and instead of defining each component of the regression model, it is equivalent to define explicitly the regression model as a polynomial of degree d with $d=\overline{1,4}$.

R code used:

```
##model1 = linear
model1=lm(y~poly(t,degree=1,raw=T),lwd=2)
summary(model1)
lines(smooth.spline(t,predict(model1)), col="maroon",lwd=2)
##model2 = quadratic
model2=lm(y~poly(t,degree=2,raw=T),lwd=2)
summary(model2)
lines(smooth.spline(t,predict(model2)), col="green",lwd=2)
```

```
##model3 = cubic
model3=lm(y~poly(t,degree=3,raw=T),lwd=2)
summary(model3)
lines(smooth.spline(t,predict(model3)), col="dodgerblue",lwd=2)
##model4 = quartic
model4=lm(y~poly(t,degree=4,raw=T),lwd=2)
summary(model4)
lines(smooth.spline(t,predict(model4)), col="red",lwd=3)
##legend
legend(1880, 0.8,legend=c("Polynomial of degree 1 ","Polynomial of degree 2" ,"Polynomial of degree 3","Polynomial of degree4") ,col=c("maroon","green", "dodgerblue" ,"red") ,lwd=2)
##plot residuals
par(mfrow = c(2, 2))
plot(model1,main="Polynomial of degree 1", which=c(1,1))
plot(model2,main="Polynomial of degree 2", which=c(1,1))
plot(model3,main="Polynomial of degree 3", which=c(1,1))
plot(model4,main="Polynomial of degree 4", which=c(1,1))
```

Numerical results

The following table summarizes the most important characteristics of each model; observations values are taken from the summary command, which gives us many numbers, but the most important values for the four models, which are the estimates coefficients and its p-value, standard error S and adjusted R-squared R_a^2 shall be recognized.

Model	**Estimates**	**p-value**	S	R_a^2
Linear	$\widehat{\beta_0} = -1.277 \times 10^1$	$< 2 \times 10^{-16}$	0.1335	0.7946
	$\widehat{\beta_1} = 6.589 \times 10^{-3}$	$< 2 \times 10^{-16}$		
Quadratic	$\widehat{\beta_0} = 2.803 \times 10^2$	$< 2 \times 10^{-16}$	0.0774	0.9313
	$\widehat{\beta_1} = -2.944 \times 10^{-1}$	$< 2 \times 10^{-16}$		
	$\widehat{\beta_2} = 7.726 \times 10^{-5}$	$< 2 \times 10^{-16}$		
Cubic	$\widehat{\beta_0} = -1.657 \times 10^3$	0.0986	0.0766	0.9332
	$\widehat{\beta_1} = 2.691$	0.0819		
	$\widehat{\beta_2} = -1.456 \times 10^{-3}$	0.0670		
	$\widehat{\beta_3} = 2.623 \times 10^{-7}$	0.0539		
Qaurtic	$\widehat{\beta_0} = 2.187 \times 10^5$	6.66×10^{-5}	0.0724	0.9409
	$\widehat{\beta_1} = -4.501 \times 10^2$	6.48×10^{-5}		
	$\widehat{\beta_2} = 3.473 \times 10^{-1}$	6.30×10^{-5}		
	$\widehat{\beta_3} = -1.191 \times 10^{-4}$	6.10×10^{-5}		
	$\widehat{\beta_4} = 1.532 \times 10^{-8}$	5.89×10^{-5}		

Table 3.2: Most important characteristics of summary of the four models

Based on the 'p-value' we can conclude that the lesser the p-value the more significant is the variable, if the p-value is less than 0.05 (the default level), the predictor is an influential factor in the variable we are trying to study. Consequently, from the 'summary' dump we notice that the estimate $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\widehat{\beta_3}$ of the cubic model are less significant features as the 'p' value is large for them, however, all the predictors are significant in linear, quadratic and quartic model.

We can assume which model represents data well based on R_a^2 value as it indicates how much does variation the model capture. R_a^2 closer to one indicates that the model explains the large value of the variance of the model, hence a good fit. In this case, the value is $R_a^2 = 0.9409$ (closer to one) of the quartic model and also it has the smallest residual standard error $S = 0.0724$. As a start, we can say that quartic model represents data well

Some diagnostic plots:

At this point, it might be useful to plot all the four models, in order to analyze the good trend and take an idea about the model that adequately approximates our data. In fact, R does not have a function for plotting polynomials model, which are found. Therefore, we must use the function 'predict' which calculates the y values given the x values; the coordinates are linked with lines. The previous code gives this plot:

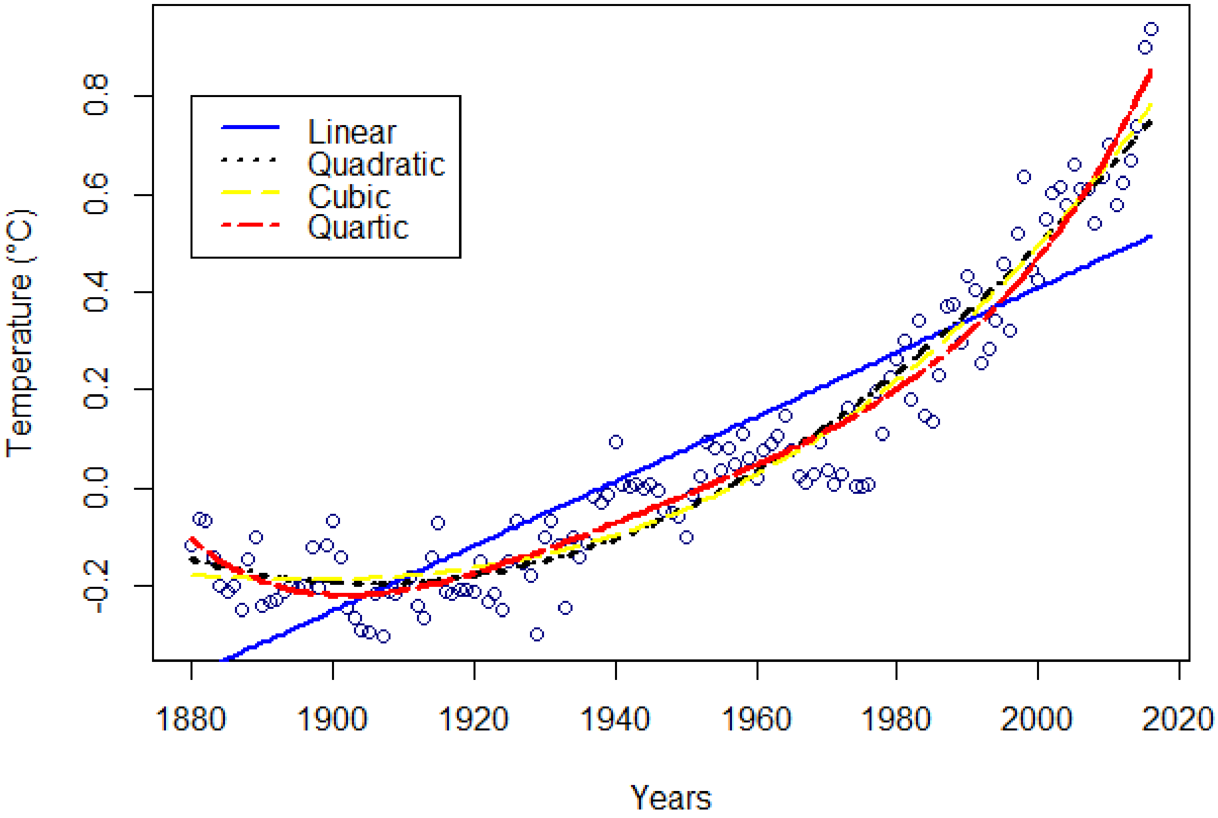

Figure 3.2: The curve fitting for the four models

Based on this graphic, we note that linear model does not describe the global trend in data, in contrast, with the rest of the models that fit the shape of the general trend in a good way. We also note that the quadratic and cubic models have very similar trends.

The residual plots:

It is important to verify some assumptions in which your data must meet in order to have valid results. In this part, we focus on the assumptions that the residuals terms have a mean of zero and constant variance. As so as to review these assumptions, we should use residuals versus fitted values plot, which show the difference between the observed response and the fitted response values. The ideal residual plot, called the null residual plot, shows a random scatter of points forming an approximately constant width band around the identity line. If, for example, the residuals increase or decrease with the fitted values in a pattern, the errors may not have a constant variance. The following figure given as the residual plot for each model:

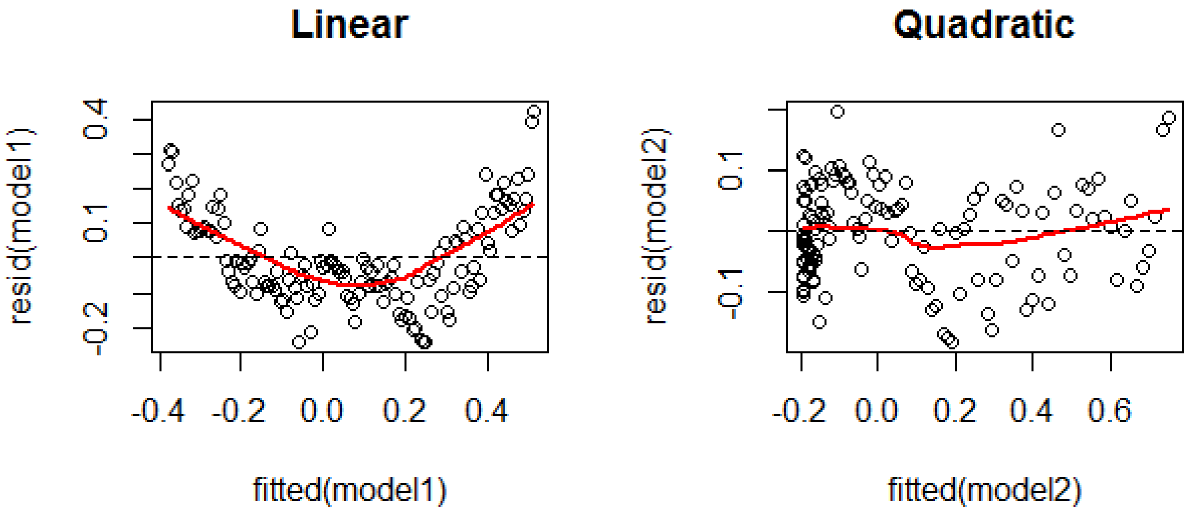

Figure 3.3: Residuals versus fitted values plot for linear and quadratic models

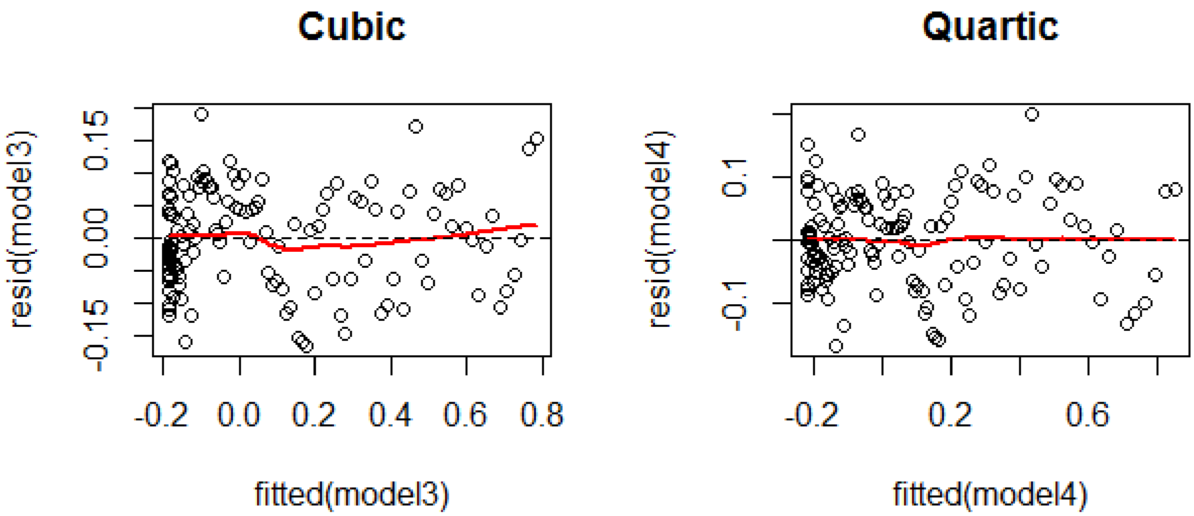

Figure 3.4: Residuals versus fitted values plot for cubic and quartic models

For linear model, there is definitely a noticeable pattern here. Residuals take on positive values with small or large fitted values, and negative values in the middle. The points are not randomly scattered around the zero line from left to right. This graph tells us we should not use the regression model that produced these results.

Observing the other plots, there are no obvious patterns, but for quadratic and cubic model, the residuals show a slight (increasing and decreasing) trend which suggests that the residuals are not identically distributed around zero. However, the residuals' trend of the quartic model began to disappear and it does not appear to increase or decrease across the fitted values, so, we can assume that the variance in the error terms is constant. The points on the plot above appear to be randomly scattered around zero, therefore, we assume that the error terms have a mean of zero is reasonable.

3.3.2 Model selection

The sum of data analyzes and the available models being studied, on the other hand, model selection procedure can help identifying the most appropriate model that fit our data. In fact, there are several quantitative criteria and several statistical tests, which are available for comparing models, thus, we present two of them.

Analyses of variance of two models

We compare two nested models: a full model y_2 and a reduced model y_1, we have:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon \\ y_2 &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \beta_{k+1} x_{k+1} + ... + \beta_p x_p + \varepsilon \end{aligned}$$

We intend to test the hypothesis in which the full model adds explanatory value over the reduced model. That hypothesis is:

$$H_0 : \beta_{k+1} = ... = \beta_p = 0$$

The statistic we use is the following:

$$\begin{aligned} F &= \frac{\text{explained variance}}{\text{unexplained vriance}} \\ &= \frac{(SSE_{y_1} - SSE_{y_2})/(p-k)}{SSE_{y_2}/(n-p-1)} \end{aligned}$$

with, SSE_{y_1}, SSE_{y_2} is the Error sum of squares for models y_1, y_2 respectively[1].

Under the null hypothesis, the F test follows a Fisher distribution with $(k-p, n-k-1)$ degree of freedom thus, we can compute a p-value associated.

The F statistic test, if the model includes more predictors (full model) it will be significant better than the reduced model, it has two hypothesis; the null hypothesis clarifies that there is no significant difference between the two models, while the alternative hypothesis states that the full model is more significant.

The relevant values are produced by the ANOVA function in R. This function compares

[1] See:www.calvin.edu/~stob/courses/m241/F11/.../Nov29-anova.pdf.

a reduced model to a full model. the following table gives a brief of the anova test:

Test	F	**p-value**
anova(linear,quadratic)	266.55	$< 2.2 \times 10^{-16}$
anova(quadratic,cubic)	3.7829	0.05389
anova(cubic,quartic)	17.235	5.888×10^{-5}

Table 3.3: Result of ANOVA

As stated in table (3.3), and based on the p-value of the first test between linear and quadratic model that is less than 0.05, we reject the null hypothesis and conclude that there is a reliable evidence considers the quadratic model statistically significant and better fits than the linear model.

For the second test, we can notice that the p-value is large, p=0.05389, thus, there is not a statistically significant difference in the two models hence, we can decide that the cubic term does not improve the model. However, we note that the quartic model is slightly preferred to cubic model with p-value equal to 5.888×10^{-5}.

Information cretiria

Information criteria such as the Akaike information criterion (AIC) and the Bayesian information criaterion (BIC) can also be used for comparing models. AIC and BIC are both penalized-likelihood criteria,defined by:

$$AIC = -2\ln(L) + 2p,\ BIC = -2\ln(L) + \log(n)p$$

where p is the number of parameters to estimate the model and L is the maximum of the likelihood function of the model. The objective of these criteria is to propose a model with an optimal compromise between the goodness of fit (measured by the log-likelihood) and the complexity of the model (measured by the number of parameters p). The best model is generally the one that minimizes both AIC and BIC. Therefore, we calculate it for the four models in R with the following command: AIC(model1,model2,model3,model4), BIC(model1,model2,model3,model4)

Result:

Test	**Linear**	**Quadratic**	**Cubic**	**Quartic**
AIC	-159.0447	-307.0581	-308.9004	-323.7131
BIC	-150.2848	-295.3782	-294.3005	-306.1932

Table 3.4: Result of AIC and BIC for linear, quadratic, cubic and quartic model

Models with the lowest AIC and/or BIC are preferred. Here, both criteria agree on rejecting Linear model with high confidence and the quartic model have the lowest values of AIC and BIC, which is considered the best model for this criterion.

Chosen model

The following table represents all the important values of the four models of the previous analysis.

Model	S	R_a^2	AIC	BIC
Linear	0.1335	0.7946	-159.0447	-150.2848
Quadratic	0.0774	0.9313	-307.0581	-295.3782
Cubic	0.0766	0.9332	-308.9004	-294.3005
Quartic	0.0724	0.9409	-323.7131	-306.1932

Table 3.5: S, adjusted R-squared, AIC and BIC result for each model

According to these values, we find that the quartic model is the most appropriate, inspecting the other fit indices, the quartic model has the smallest value of the residual standardnerror S, which is more significantly than the other models, while the two information criteria confirmed that quartic model is a good choice. Furthermore, the R_a^2 is the highest of all models. Considering the overall summed up evidence, it seems reasonable to conclude that the quartic model is the most parsimonious and best-fitting model for this data set. Finally, the fitted model of the global temperature is given by:

$$\widehat{y} = 2.187 \times 10^5 + -4.501 \times 10^2 t + 3.473 \times 10^{-1} t^2 + -1.191 \times 10^{-4} t^3 + 1.532 \times 10^{-8} t^4.$$

3.3.3 Confidence and prediction interval of the fitted model

As final step, when you fit a parameter to a model, the accuracy or precision can be expressed as a confidence interval or a prediction interval, the two are quite distinct.

R code used:

```
r=predict(model4,interval="confidence")
s=predict(model4,interval="prediction")
r1=r[,2]
r2=r[,3]
s1=s[,2]
s2=s[,3]
lines(smooth.spline(t,r1), col="darkgoldenrod1",lwd=4,lty = 3)
lines(smooth.spline(t,r2), col="deepskyblue",lwd=4,lty = 3)
lines(smooth.spline(t,s2), col="green",lwd=2)
lines(smooth.spline(t,s1), col="deeppink",lwd=2)
legend(1880, 0.8,legend=c("fitted line","Lower CI ","Upper CI","Upper PI","Lower PI"), col=c("red","darkgoldenrod1","deepskyblue","green","deeppink"),
lty=c(1,3,3,1,1),lwd=2,bty="n",cex=0.9)
```

The figure below explains each of them on the basis of the fitted model.

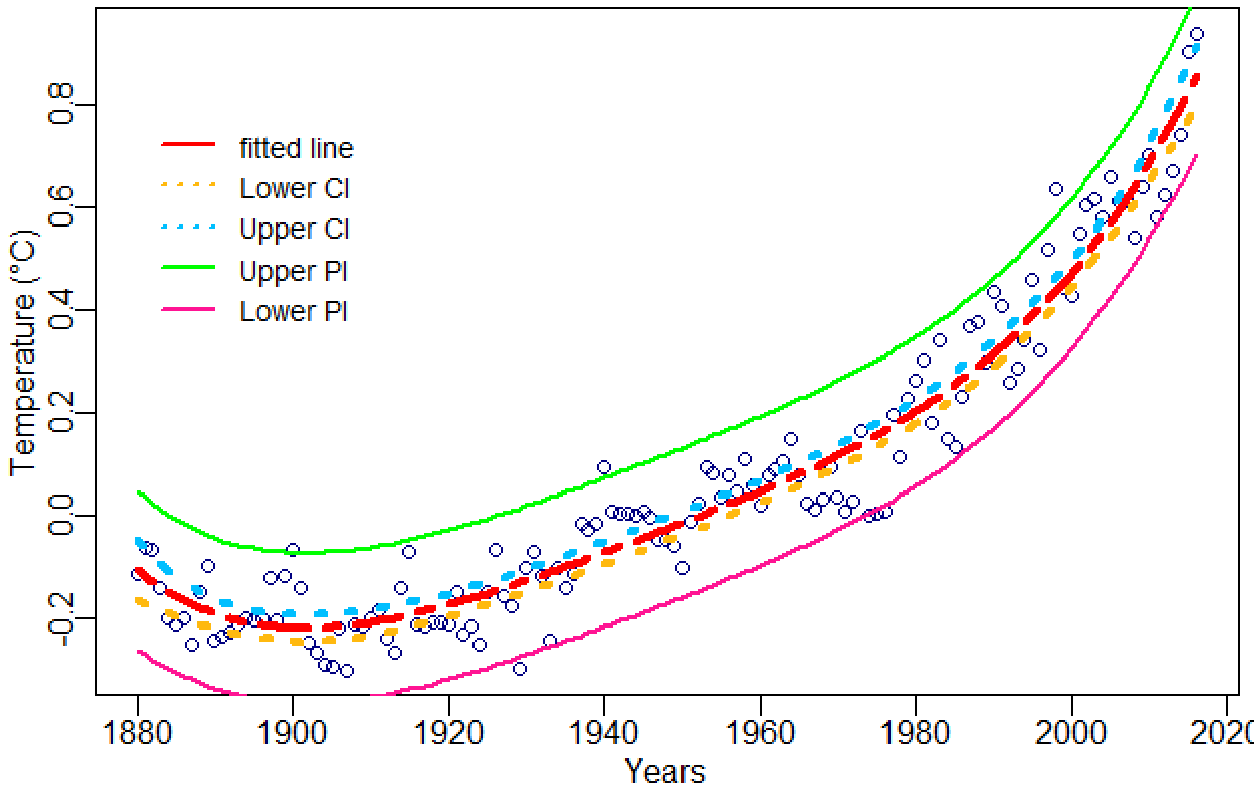

Figure 3.5: Confidence and prediction interval of the fitted model

The figure shows the confidence interval for the global temperature, which there is a 95% probability that the true best-fit line for the data lies within the two dotted lines. Moreover, for the prediction interval there is a 95% of the y-values are found for a certain x-value within the interval range around the regression line except three values that are considered outliers.

We note that the prediction interval is wider than the confidence interval of the prediction.

3.4 Conclusion

The example shows how to approach linear regression modeling for the global temperature over years. We focus our study on polynomials, which are powerful tools. In this case, we find that data was generated using a fourth degree polynomial, however, The model that is created still has a scope for improvement as we can apply techniques like Outlier detection, Correlation detection to further improve the accuracy of more accurate prediction. As a matter of fact, when analyzing real data, we usually know few about it, therefore, we need to be cautious, because the use of high order polynomials ($d > 4$) may lead to over-fitting. Even though your model is getting better at fitting the existing data, this can be bad when you try to predict new data and lead to misleading results.

Conclusion

This master's thesis explored the linear regression model, in which a dependent variable is controlled or affected by a set of independent variables. The purpose of regression analysis is to establish a relationship between response variable and predictors, also, to predict dependent variable based on a set of values of independent variables, moreover, to identify which predictor is more important than others are and to explain the response variable so that the relationship can be more efficiently and accurately.

The commonly used techniques for estimating the parameters of regression are the OLS and ML. Actually, the ML is a method used in estimating the parameters of a statistical model and for fitting a statistical model to data. It is used when the functional form of the probability distribution of the error term is specified. While, the OLS is a general method for approximately determining the unknown parameters located in a linear regression model, moreover, no matter what may be the form of the distribution of the error terms, the OLS method provides unbiased point estimator that have a minimum variance among all unbiased linear estimators. However, we need to make an assumption about the form of the distribution of the error to set up interval estimates and tests and a normal error term greatly simplifies the theory of regression analysis and it is justifiable in many real world situations.

In this project, we emphasized on presenting a specific regression techniques including simple linear regression analysis, multiple linear regression analysis, and we gave a highlight about nonlinear regression. Particularly, we concentrated on studying a special case of data that is global temperature. We first reviewed its scatter plot, which has a curved trend. Hence, in order to model this data, we proposed a polynomial regression that is to model a non-linear relationship between the independent and dependent variable, but as a statistical problem, it is considered linear. Therefore, we build typically four models to fit our data before selecting the best model. For this, we used a useful commands and diagnostics in software R; we relied on the function lm(), which fits a model using Ordinary Least Squares (OLS) method and we end this analysis by functions and parameters to support a number of criteria for selecting models in R.

We conclude that the use of polynomial regression model may lead to the known collinearity issues because of the high order polynomials. Therefore, the models developed using regression analysis are not perfect, and there is many ways to improve our regression model.

A researcher can use an advanced technique like Random Forest and Boosting technique. As a future project, we propose the application of Stochastic Gradient Boosting and Nonlinear Regression Splines that handle the missing values, interactions, outliers and nonlinearities in data.

Bibliography

[1] Abdalla, A. M. A. (2013). Joint Regression and Ordination Analysis Techniques of Gxe Interactions for the Shortening Growing Season of Cotton. Egyptian Journal of Plant Breeding, 203(1132), 1-18..

[2] Andersen, P. K., & Skovgaard, L. T. (2010). Regression with linear predictors. Springer Science & Business Media.

[3] Chatterjee, S., & Hadi, A. S. (2015). Regression analysis by example. John Wiley & Sons.

[4] David J. Olive (auth.)-Linear Regression-Springer International Publishing (2017).

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.

[6] Kass, R. E. (1990). Nonlinear regression analysis and its applications. Journal of the American Statistical Association, 85(410), 594-596.

[7] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

[8] Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models (Vol. 4). Chicago: Irwin.

[9] NOAA National Centers for Environmental information, Climate at a Glance: Global Time Series, published April 2018, retrieved on May 7, 2018 from http://www.ncdc.noaa.gov/cag/.

[10] Stine, R., & Foster, D. (2017). Statistics for Business: Decision Making and. Addison-Wesley SOFTWARE-JMP.

[11] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.

[12] Yan, X., & Su, X. (2009). Linear regression analysis: theory and computing. World Scientific.

Annex A: softwares: R and SPSS

SPSS:

Statistical Package for Social Sciences (SPSS) is also one of the most widely used softwares for the statistical analysis in the area of social sciences. It is one of the preferred softwares used by market researchers, health researchers, survey companies, government, education researchers, among others. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored) are features of the SPSS. Many features of SPSS are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language.

R:

R is a language and environment for statistical computing and graphics. It is a GNU project similar to the S language and environment. R can be considered as a different implementation of S language. There are some important differences, but much code written for S language runs unaltered under R. The S language is often the vehicle of choice for research in statistical methodology, and R provides an open source route to participation in that activity. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. R is available as free software under the terms of the Free Software Foundation's GNU General Public License in source code form.

Annex B: Abreviations and Notations

$i = \overline{1, n}$	the index i is incremented by one unit from 0 to n
$\overline{x}$	Sample mean
S^2	Sample Variance
S	Sample standard deviation
$Var(.)$	Variance.
$E(.)$	Mathematical expectation (Mean)
$Cov(.)$	Covariance
$N(0, \sigma^2)$	Normal distribution with mean 0 and varince σ^2
$N_n(0, \sigma^2)$	Multivariate normal distribution with mean 0 and variance σ^2
A^t	Transpose of matrix A
I_n	Identity matrix size $(n \times n)$
$0_{n\times 1}$	Null vector
$X \ \rightsquigarrow$	Distribution of X
$\mathbb{R}$	Real numbers
$\prod_{i=1}^{i=n}$	Product from $i = 1$ to $i = n$
$\sum_{i=1}^{i=n}$	Sum from $i = 1$ to $i = n$

$\log(x)$	Commun Logarithm
$\ln(x)$	Natural Logarithm
Df	Degree of freedom
$p - value$	Probability value