National Institute of Science Education and Research

School of Mathematical Sciences

M498: Project Report

Theory of Regression Analysis with Applications

Student
Tanikella Padma Ragaleena
4th year Integrated M.Sc.
School of Mathematical Sciences
NISER Bhubaneswar (HBNI)
tp.ragaleena@niser.ac.in

Supervisor
Dr. Shyamal Krishna De
Reader-F
School of Mathematical Sciences
NISER Bhubaneswar (HBNI)
sde@niser.ac.in

November 21, 2019

Acknowledgements

I would like to express my sincere gratitude to my supervisor *Dr. Shyamal Krishna De* for his patience, motivation, and immense knowledge.

Contents

Chapter 1

Linear Regression

Linear Regression is a branch of statistics that aims to examine relationships between one or more variables to create models that can be used for predictive purposes. It was first invented by the English scientist Sir Francis Galton while he was studying the relationship between heights of fathers and heights of their first sons. He wanted to predict the height of son based on the height of his father. He fitted a line to the scatterplot of these heights because it showed an increasing and linear trend. This fitted line depicted that given a father's height, the height of his son will show a regression towards the mean. (The word regression in simple English means a return to former or less advanced/worse state/condition/way of behaving.) This is how the procedure of fitting a linear equation to given data started being called **linear regression**.

The sources used to make first chapter were [16] and [15].

1.1 Introduction

The objective of this chapter is to model the data we have at hand so that we can use the model **to predict** outcomes. Regression analysis is the process of finding a mathematical equation that best fits the data given to us. Note that the use of the techniques in Linear regression implies that we are assuming the existence of a linear relationship between the variables at hand. For this chapter, we assume that there exists one independent variable and one or more than one dependent variables.

Our first step in regression analysis should be to hypothesize the model that fits our data to be

$$\boxed{Y = f(x_1, x_2, \cdots, x_k; \beta_0, \beta_1, \beta_2, \cdots, \beta_k) + \epsilon}$$

where Y denotes the dependent variable which is a **random variable** and the x_i for $i \in \{1, 2, \cdots, k\}$ are the independent variables which are **non-random**. ϵ denotes the error random variable which has zero mean and constant variance. β_j for $j \in \{0, 1, 2, \cdots k\}$ denote the unknown parameters that need to be predicted. The dependent variable Y is also called the *response variable* and the independent variables x_i are also known as the *predictor variables*.

In case of **Multiple Linear Regression**, the model assumed for the given data is

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon \text{ where } E(\epsilon) = 0 \tag{1.1}$$

We assume that all x_i are non-random and known without any measurement errors. So since $E(\epsilon) = 0$, we have that

$$\begin{aligned} &E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\ &\Rightarrow\ Y = E(Y) + \epsilon \end{aligned}$$

Hence Y has a deterministic component $E(Y)$ and an error component ϵ. ϵ expresses the inability to provide an exact model for a given natural phenomenon.

Remark 1.1. 1. The case where we assume one dependent and one independent variable is called **simple linear regression**.

2. Before applying simple linear regression, we can identify the need to use it using a **scatter-plot** which helps us to see the approximate linear trend in data. Another way to check for the amount of linear relationship is data is through **Pearson's correlation coefficient**. Scatter-plot gives a visual representation of the linear trend while the correlation coefficient quantifies the amount of linear relationship.

3. We assume the expectation of the error to be 0 without loss of generality because a model with non-zero error can always be modified to get a model with 0 mean for error. For simplicity consider the case of simple linear regression

$$\text{Model: } Y = \beta_0 + \beta_1 x + \epsilon \text{ such that } E(\epsilon) = k \neq 0$$

$$\text{So consider } Y = \beta_0^* + \beta_1 x + \epsilon^* \text{ where } \beta_0^* = \beta_0 + k \text{ and } \epsilon^* = \epsilon - 3 \text{ such that } E(\epsilon^*) = 0$$

1.2 Method of Least Squares

In order to predict the linear model Equation 1.1, we first need to predict the unknown parameters β_i for $i \in \{1, 2, \cdots k\}$. Doing this will let us predict the dependent variable Y by substituting the known independent variable values in the Equation 1.1. Hence a linear model to fit the the given data is obtained by estimating the values of unknown parameters.

There are a number of ways to fit a line to the given data. One of the possible options is **the method of least squares**. First we consider the case of simple linear regression. Consider "n" bivariate data points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ with errors $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ respectively. So we have the following n equations :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ where } i \in \{1, 2, \cdots, n\}$$

We assume here that $\epsilon_i's$ are independent and identically distributed normal random variables with mean 0 and variance σ^2. The parameters are estimated by minimizing a quantity known as **sum of squared errors** which is given by :

$$\text{SSE} = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

In the method of least squares, we find the values of β_0 and β_1 which minimize the sum of squared errors. These values which are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the **least square estimates** of the parameters β_0 and β_1. Also, the line $Y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the **least squares line** which can be used to predict the Y values for given x values.

Remark 1.2. The predicted value of y_i, denoted by $\hat{y}_i$, is given by $\hat{\beta}_0 + \hat{\beta}_1 x_i$. The quantity $y_i - \hat{y}_i$ is called the i^{th} **residual** and it is denoted by e_i. Hence

$$\text{SSE } = \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{i=n} e_i^2$$

So minimizing SSE is the same as finding the parameter values which minimize the deviation of y_i from $\hat{y}_i$.

The following result from several variable calculus will be used to derive the least square estimators.

Theorem 1.1. *Let $f(x, y)$ be function in two variables with the Hessian matrix H. Let its determinant be denoted by D and let (a, b) be a critical point for f i.e. $\nabla f(a, b) = 0$. Then (a, b) is a relative minima for f if $D > 0$ and $h_1 1 > 0$. Note that here, we evaluate D and H at the point (a, b) and $H = (h_{ij})$.*

Theorem 1.2.

$$\text{Let } S_{xx} = \sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{n}$$

$$\text{and } S_{xy} = \sum_{i=1}^{i=n} x_i y_i - \frac{\left(\sum_{i=1}^{i=n} x_i\right)\left(\sum_{i=1}^{i=n} y_i\right)}{n}$$

Then $\boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}}$ *and* $\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$ *where $\bar{x}$ and $\bar{y}$ denote mean of all corresponding data points.*

Proof.

$$\text{Let } f(\beta_0, \beta_1) = \text{ SSE } = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Finding critical points i.e. we solve $\nabla f = 0 \implies \left(\frac{\partial f}{\partial \beta_0}, \frac{\partial f}{\partial \beta_1}\right) = (0, 0)$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = 0 \text{ gives } 2\left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i\right) = 0 \tag{1.2}$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = 0 \text{ gives } (-2)\left(\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2\right) = 0 \tag{1.3}$$

(1.2) and (1.3) together are called **Least Square Equations** which on rearrangement give the following **Normal Equations**

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \text{ and } \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \tag{1.4}$$

Solving the normal equations for β_0 and β_1 gives us the expression for least square estimators as $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Hence our critical point is $(\hat{\beta}_0, \hat{\beta}_1)$ which we should check if it is a maxima or a minima.

$$\text{We have D } = f_{xx} f_{yy} - f_{xy}^2 = 2n \times 2\sum_{i=1}^{n} x_i^2 - \left(2\sum_{i=1}^{n} x_i\right)^2 = 4n^2\sigma^2 > 0$$

$$\text{Also } \frac{\partial^2 f}{\partial {\beta_0}^2} = 2n > 0$$

Hence $(\hat{\beta}_0, \hat{\beta}_1)$ is a point of relative minima. In fact its a point of global minima since f is a convex function. □

Remark 1.3. It is important to observe that the least squares method did not make use of the fact that the error terms are IID normal random variables.

Theorem 1.3. *$SST = SSE + SSReg$*
where $SST =$ total sum of squares $= \sum_{i=1}^{n}(y_i - \bar{y})^2$
SSE or $SSRes =$ sum of squares of error $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
and $SSReg =$ sum of squares of regression $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Proof.

$$SST = \text{ total sum of squares } = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right)^2$$

Expanding the above term in RHS gives SSReg and SSE on the RHS. □

NOTE:

1. SST measures the amount of variation of $y_i's$ around $\bar{y}$ i.e. it is a measure of amount of variation in the data.
2. SSE measures the lack of fit in the regression model.
3. SSReg is a measure of the variation that can be described by the regression model.

1.2.1 Quality of Regression

Assume that we were given a set of data points for which we fit a line which is best in some sense. Then the next question that we should ask is "How well does the line fit the data?". It would be nice to come up with a formula using which we can quantify the quality of fit. This quantity is called the **Coefficient of Determination** (r^2). It is given by

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Observe that the quantity $\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ gives a measure of the proportion of variation that is *not* described by the model. Since $SST = SSE + SSReg$, we can say that the quantity $1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ describes the amount of variation described by the model.

1.2.2 Properties of the least square estimators

First of all, we should know that knowing the distributional properties of β_0 and β_1 is useful because it helps us make statistical inferences about the parameters and hence ultimately about the model. In the properties that will be listed in this section, we will assume a simple linear regression model i.e. $Y_i = \beta_0 + \beta_1 x_i + \epsilon$ where ϵ_i $N(0, \sigma^2)$ for $i \in \{1, 2, \cdots, n\}$ and ϵ_i and ϵ_j are independent of each other for all $i \neq j$.

Theorem 1.4. *The least square estimators follow the following two properties :*

1. *$\hat{\beta}_0$ and $\hat{\beta}_1$ follow a normal distribution.*
2. *$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1 respectively.*

Proof.

$$\text{Note that } S_{xx} = \sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{n} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1.5}$$

$$S_{xy} = \sum_{i=1}^{i=n} x_i y_i - \frac{\left(\sum_{i=1}^{i=n} x_i\right)\left(\sum_{i=1}^{i=n} y_i\right)}{n} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{1.6}$$

Using Equations 1.5 and 1.6 , we can write $\hat{\beta}_1$ as follows by expanding the numerator S_{xx}

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{S_{xx}}$$

Since all the $x_i's$ are known without measurement errors, so we can say from above equation that $\hat{\beta}_1$ follows a normal distribution as $\hat{\beta}_1$ has been written as a linear combination of normal random variables (Y_i).

We know that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ where $\bar{y}$ and $\hat{\beta}_1$ are normal random variables because they are linear combinations of known normal random variables. Hence $\hat{\beta}_0$ also follows normal distribution.

Proof that least square estimators are unbiased:

$$E[\hat{\beta}_1] = \frac{\sum_{i=1} n(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} = \frac{\beta_1}{S_{xx}} \times \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \frac{\beta_1}{S_{xx}} \times S_{xx} = \beta_1 \tag{1.7}$$

$$E[\hat{\beta}_0] = E[\bar{Y} - \hat{\beta}_1 \bar{x}] = E[\bar{Y}] - \bar{x} E[\hat{\beta}_1] = \frac{\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)}{n} - \beta_1 \bar{x} = \beta_0 \tag{1.8}$$

The Equations 1.7 and 1.8 show that the least square estimators are both unbiased estimators. □

Remark 1.4. It is to be noted that we have used the fact that errors are IID normal random variables to understand the distributional properties of estimators but the least squares method had nothing to do with the kind of distribution for errors.

Definition 1.1. An estimator $\hat{\theta}$ is called the **Best Linear Unbiased Estimator (BLUE)** for θ if

1. $\hat{theta}$ is a linear combination of sample observations
2. $Var(\hat{\theta}) \leq Var(\hat{\theta}')$ where $\hat{\theta}'$ is any other estimator which is unbiased.

So we first restrict our set of estimators to those which are linear in data and unbiased. Among all these estimators, we pick the one with minimum variance and call it BLUE

Theorem 1.5. ***: Gauss-Markov Theorem***
Let $Y = \beta_0 + \beta_1 x + \epsilon$ be a simple linear regression model such that $Y = (y_1 \cdots y_n)^T$ and $\epsilon = (\epsilon_1 \cdots \epsilon_n)$ with all $\epsilon_i's$ to be IID random variables. Then the least square estimators for β_0 and β_1 are Best Linear Unbiased Estimators(BLUE).

Proof. We will see the proof of a more general case of Gauss-Markov theorem later. □

Theorem 1.6. *Few useful properties of the least square estimators are:*

1. $cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{S_{xx}}$
2. $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *are independent of* SSE

Proof. 1.

$$\text{We know that } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}} = \sum_{i=1}^{n} c_i y_i \text{ where } c_i = \frac{x_i - \bar{x}}{S_{xx}} \tag{1.9}$$

$$\text{Now, } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x} \right) y_i = \sum_{i=1}^{n} d_i y_i \text{ where } d_i = \frac{1}{n} - c_i \bar{x} \tag{1.10}$$

Note that we have used Equation 1.9 to arrive at the Equation 1.10. Now the covariance is

$$cov\left(\sum_{i=1}^{n} c_i y_i, \sum_{i=1}^{n} d_i y_i \right) = \sigma^2 \sum_{i=1}^{n} c_i d_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} \left(\frac{1}{n} - \frac{x_i - \bar{x}}{S_{xx}} \bar{x} \right) = \frac{-\sigma^2 \bar{x}}{S_{xx}}$$

2. We use the following lemma to prove this

Lemma 1.1. *If $X_1 X_2 \cdots X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i^2)$ and $U = \sum_{i=1}^{n} a_i X_i$ and $V = \sum_{i=1}^{n} b_i X_i$ where a_i, b_i are constants, then*

$$cov(U, V) = 0 \text{ iff } U \text{ and } V \text{ are independent} \tag{1.11}$$

$$SSE = \sum_{i=1}^{n} e_i^2 \text{ and } \beta_1 = \sum_{i=0}^{n} c_i y_i, \beta_0 = \sum_{i=0}^{n} d_i y_i$$

Clearly, it is enough to show that $cov(\hat{\beta}_1, e_i) = 0$ and $cov(\hat{\beta}_0, e_i) = 0$ because of the lemma above. So

$$cov(\hat{\beta}_1, e_i) = cov(\hat{\beta}_1, y_i - \hat{\beta}_0 x_i) = cov(\sum_{i=1}^{n} c_i y_i, y_i) - cov(\hat{\beta}_0, \hat{\beta}_1) - Var(\hat{\beta}_1 x_i)$$

Hence we have,

$$cov(\hat{\beta}_1, e_i) = \sigma^2 c_i + \frac{\sigma^2 \bar{x}}{S_{xx}} - \frac{\sigma^2 x_i}{S_{xx}} = \frac{\sigma^2}{S_{xx}} [x_i - \bar{x} + \bar{x} - x_i] = 0$$

Similarly we can show that $cov(\hat{\beta}_0, e_i) = 0$. This proves the result mentioned in theorem. □

Theorem 1.7. *The least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are Maximum Likelihood Estimators (MLE) of β_0 and β_1 respectively.*

Proof. The proof of the same statement is given in the more general multi-variable case. □

1.3 Multiple Linear Regression

Earlier, when we were studying simple linear regression, we considered single regressor variable, but in multiple linear regression we assume more than one regressor variable. Hence the model we have here is

$$Y = \beta_0 + \beta_1 x_1 \cdots \beta_k x_k + \epsilon \text{ where } \beta_0 \cdots \beta_k \text{ are regression coefficients}$$

Just like in simple linear regression, here also Y is a random variable, $\beta_i' s$ are unknown parameters that need to be estimated and $x_i' s$ are the known predictor variables.

Remark 1.5. Note that in linear regression, linear refers to being linear with respect to the parameters and not the predictors. So

$$Y = \beta_0 + \beta_1 x \cdots \beta_k x^k + \epsilon \text{ is a linear model while } Y = \beta_0 + \beta_1^2 x_1 \cdots \beta_k^3 x_k + \epsilon \text{ is not}$$

1.3.1 Matrix representations

Consider the model

$$Y = \beta_0 + \beta_1 x_1 \cdots \beta_k x_k + \epsilon \tag{1.12}$$

where $Y = (y_1, y_2, \cdots, y_n)^T$ and the predictor variables with respect to y_i are $(x_{i1}, x_{i2}, \cdots, x_{ik})$. This information has been summarized in the table below:

Obs. num.	Response	Predictors
1	y_1	$x_{11}, x_{12}, x_{13}, \cdots, x_{1k}$
2	y_2	$x_{21}, x_{22}, x_{23}, \cdots, x_{2k}$
.	.	$\cdots$
.	.	$\cdots$
n	y_n	$x_{n1}, x_{n2}, x_{n3}, \cdots, x_{nk}$

Table 1.1: Model

Considering the n observations, our model 1.12 can be split into the following n equations

$$y_i = \beta_0 + \beta_1 x_{1i} \cdots \beta_k x_{ki} + \epsilon_i \text{ for } i \in \{1, 2, \cdots n\} \tag{1.13}$$

It is easy to see that the all the information in the model can be compressed in a matrix equation as given below:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In short this is written as $y = X\beta + \epsilon$ where X is the $n \times p$ matrix of predictors and it is called the **Design matrix**. Just like in the simple linear regression case, here also we make certain assumptions about our model, which are: $E(\epsilon) = 0, Var(\epsilon) = \sigma^2 I_n$ and $\epsilon \sim N(0, \sigma^2 I)$

It is useful to know the multivariate normal distribution which is as follows:

$$\text{If } X \sim N(\tilde{\mu}, \sum) \text{ then } f(\tilde{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\sum|}} \exp\left(\frac{1}{2}(\tilde{x} - \tilde{\mu})^T \Sigma^{-1} (\tilde{x} - \tilde{\mu})\right)$$

where the Σ is a non-singular matrix and X is a full-rank matrix.

Some useful results

Here **a** is a $k \times 1$ vector of constants, **A** is a $k \times k$ matrix of constants and **y** is a $k \times 1$ random vector with mean μ and non-singular variance-covariance matrix **V**.

- $\mathrm{E}(a'y) = a'\mu$
- $\mathrm{E}(Ay) = A\mu$
- $\mathrm{Var}(a'y) = a'Va$
- $\mathrm{Var}(Ay) = AVA'$
- $\mathrm{E}(y'Ay) = \mathrm{trace}(AV) + \mu' A\mu$

1.3.2 Ordinary Least Square Estimation

There are two ways of arriving at the least squares estimates in the multi-variable case, one method is the usual method of minimizing the SSE and the other method has a more geometric approach. I would like to give the geometric proof below in this subsection.

Since we want to find the least square estimator in the multi-variable case, we consider the n-dimensional space. But for better visualization it is better to imagine the 3-dimensional case as well.

Let $y^T = (y_1, y_2, \cdots, y_n)$ be a vector starting from the origin in the n-dimensional space. Call the end of this vector A. The matrix X has "p" columns given by $\tilde{1}, \tilde{x_1}, \cdots, \tilde{x_k}$. Each column represents a vector from the origin and the space spanned by these columns is called the **estimation space**.

The figure below represents the case when $p = 2$ i.e. the case of simple linear regression. Since the matrix X is a full rank matrix, we can say that any point in space can be written as a linear combination of columns in X. Hence any vector in the space can be written as $X\beta$. Let the end of any such vector be called B. Then the squared distance from A to B is given by $S(\beta) = (\tilde{y} - X\tilde{\beta})^T(\tilde{y} - X\tilde{\beta})$. We also know that

$$SSE = \sum_{i=1}^{n} e_i^2 = \tilde{\epsilon^T}\tilde{\epsilon} = (\tilde{y} - X\tilde{\beta})^T(\tilde{y} - X\tilde{\beta})$$

Hence the function we want to minimize, i.e. $S(\beta)$ is the distance between points A and B. From the picture, it is clear that this distance can be minimized if B is the foot of the perpendicular when a line perpendicular to plane is dropped from A.

If B was the foot of the perpendicular, then $y - \hat{y} = y - X\beta$ should be perpendicular to the estimation space which is the x-y plane in the $p = 2$ case. Hence we get

$$X^T(y - X\hat{\beta}) = 0 \implies X^TX\hat{\beta} = X^Ty \implies \hat{\beta} = (X^TX)^{-1}X^Ty \tag{1.14}$$

Hence we have found out the parameter estimates in the multi-variable case.

1.3.3 Properties of least square estimates - multi variable case

Theorem 1.8. *The estimator $\hat{\beta}$ in Equation 1.14 is an unbiased estimator for β and $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$*

Proof.

$$E(\hat{\beta}) = E[(X'X)^{-1}X'y] = E[(X'X)^{-1}X'(X\beta + \epsilon)] = E[(X'X)^{-1}X'X\beta] + E[(X'X)^{-1}X'\epsilon] = \hat{\beta}$$

The above line is true because X is a non-random matrix and $E(\epsilon) = 0$.

Using $Var(Ay) = AVar(y)A'$ where A is a $n \times p$ matrix and y is a $p \times 1$ column vector, we get the required result that $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$.

□

Theorem 1.9. *The **Gauss-Markov** theorem establishes that the ordinary least squares estimator of β is Best Linear Unbiased Estimator (BLUE)*

Proof. We saw that "best" in single variable case meant smallest variance. Here in the multivariate case $Var(\hat{\beta})$ is a vector, so out definition of "best" has to be something else.

We consider that $\hat{\beta}$ to be the best which minimizes the variance for any linear combination of the estimated coefficients , $l'\hat{\beta}$

Now, $Var(l'\hat{\beta}) = \sigma^2 l'(X'X)^{-1}l$ = scalar. Let $\tilde{\beta}$ be another unbiased estimator of β that is a linear combination of data. Our goal, then is to show that $Var(l'\tilde{\beta}) \geq \sigma^2 l'(X'X)^{-1}l$. We use

the fact that any other estimator of β can be written as $\tilde{\beta} = [(X'X)^{-1}X' + B]y + b_0$ where B is $p \times n$ matrix and b_0 is $p \times 1$ vector of constants.

It can be shown that $E(\tilde{\beta}) = \beta + BX\beta + \beta_0$. But $\tilde{\beta}$ is assumed unbiased, hence $b_0 = 0$ and $BX = 0$. Similarly, it can be shown that $Var(\tilde{\beta}) = \sigma^2[(X'X)^{-1}X' + BB']$. Now

$$Var(l'\tilde{\beta}) = l'Var(\tilde{\beta})l = Var(l'\hat{\beta}) + \sigma^2 l'BB'l$$

Since BB' is positive semi-definite, we can see that $l'BB'l = l*'\, l* \geq 0$ where $l* = B'l$. Hence $\hat{\beta}$ is the best estimator □

Theorem 1.10. *Maximum Likelihood Estimators for the model parameters are the least square estimators. (Assuming that the model errors are independently and identically distribution).*

Proof. We know that

$$f_{\epsilon_i}(x|\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \forall i \in \{1, 2, \cdots n\}$$

$$\text{Hence, joint pdf } = f(x_1 \cdots x_n|\sigma) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} \exp\left(\frac{-\sum_{i=1}^n x_i^2}{2\sigma^2}\right) = f(\bar{\epsilon}|\sigma)$$

Hence the likelihood function is

$$L(\sigma|\bar{\epsilon}) = \frac{1}{\sigma^n\sqrt{2\pi}^n} \exp\left(\frac{-\sum_{i=1}^n \epsilon^2}{2\sigma^2}\right) = \frac{1}{\sigma^n\sqrt{2\pi}^n} \exp\left(m\frac{-\bar{\epsilon}'\bar{\epsilon}_i}{2\sigma^2}\right)$$

Writing $\bar{\epsilon}'\bar{\epsilon}_i = (\bar{y} - X\bar{\beta})'(\bar{y} - X\bar{\beta})$, we can see that likelihood is minimized w.r.t. β when $\bar{\epsilon}'\bar{\epsilon}_i$ is minimized. Since this is exactly what we minimized in the OLS case, we conclude that

$$\boxed{\hat{\beta} = MLE = \text{ least square estimates}}$$

□

1.4 Tests of Significance

In this section we derive various tests to answer questions about which x variables contribute significantly, if there is an over all linear relationship etc.

Before we formally state the tests as theorems, we will assume some statements are true.

Proposition 1.1. *A is a $k \times k$ matrix of constants. $U = y'Ay$ where $y \sim N_n(\mu, \sigma^2 I)$. If A is an idempotent matrix with rank "p", then $\frac{U}{\sigma^2} \sim \chi_p^2$.*

Proposition 1.2. *Let X be an $n \times p$ matrix partitioned such that $X = [X_1, X_2]$. Then $X(X'X)^{-1}X'X_1 = X_1'$ and $X_2'X(X'X)^{-1}X' = X_2'$*

Proposition 1.3. *Let B be a $q \times k$ matrix, and let W be a linear form given by $W = By$. The quadratic form $U = y'Ay$ and W are independent if $BVA = 0$*

Theorem 1.11. *Residual sum of squares $= SSRes$ has $n - p$ degrees of freedom associated with it. Also,*

$$\frac{SSRes}{\sigma^2} \sim \chi_{n-p}^2 \tag{1.15}$$

Proof.

$$\begin{aligned} SSRes &= (y-\hat{y})'(y-\hat{y}) \\ &= (y - X(X'X)^{-1}X'y)^T(y - X(X'X)^{-1}X'y) \\ &= y'(I - X(X'X)^{-1}X')y \end{aligned}$$

One can show that the matrix $[I - X(X'X)^{-1}X']$ is symmetric and idempotent. Also the rank of this matrix which is given by its trace, (since its a symmetric matrix) is $n-p$. Hence the theorem is true by 1.1. □

Lemma 1.2. $\frac{SSRes}{n-p}$ *is an unbiased estimator for* σ^2

Proof. We use the fact that $E(y'Ay) = trace(AV)+\mu'A\mu$ where $y \sim N_n(\mu, V)$ and that $SSRes = y'(I-X(X'X)^{-1}X')y$ (as shown in 1.11) to prove the result. We get $E(SSRes) = n-p$. Hence the lemma can be proved. □

Remark 1.6. The quantity $SSRes/n-p$ is called the **Mean Square Residual**

Theorem 1.12. $\frac{SSR}{\sigma^2} \sim \chi^2_k$ *assuming that the errors in the model are with mean 0 and constant variance of* $\sigma^2 I$.

Proof.

$$SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 \text{ where } \bar{y} = \frac{\tilde{1}'\tilde{y}}{n}$$

where $\tilde{1}$ is $n \times 1$ vector of all ones and $\tilde{y} = (y_1, \cdots, y_n)$. Also note that $n = \tilde{1}'\tilde{1}$. Using these, we get the following

$$SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 = \tilde{y}'[X(X'X)^{-1}X' - \tilde{1}(\tilde{1}'\tilde{1})^{-1}\tilde{1}']\tilde{y}$$

Now, $X = [1X_R]$ where X_R is the matrix formed by the actual values for the regressors. Hence by 1.2 , we have $X(X'X)^{-1}X'\tilde{1} = \tilde{1}$ and $\tilde{1}'X(X'X)^{-1}X' = \tilde{1}'$ Using above facts it can be shown that

$$\frac{SSR}{\sigma^2} = \frac{1}{\sigma^2}\tilde{y}'[X(X'X)^{-1}X' - \tilde{1}(\tilde{1}'\tilde{1})^{-1}\tilde{1}']\tilde{y} = \text{ call it } \tilde{y}'U\tilde{y}$$

It can be shown that U is idempotent. Hence by 1.1, $\tilde{y}'U\tilde{y}$ follows a chi-square distribution. And its rank is its trace which is equal to k. Hence statement is proved. □

1.4.1 F-test to check linearity

Now we have information to derive a hypothesis test. The first test determines if there is a linear relationship between y and $x_1 \cdots x_k$ by testing the hypothesis $H_0 : \beta_0 = \beta_1 = \cdots \beta_k = 0$ against the null hypothesis that there is at least one non-zero β_i. By bbbb, it is easy to verify that both are independent. As a result

$$\frac{MSR}{MSRes} \sim F_{k,n-p}$$

Now we can use the p-value approach or critical value approach to reject/accept the null hypothesis.

1.4.2 Tests on individual regression coefficients

Once we reject the null hypothesis of previous setting, then our next question is to know which regressors are significant. So our aim now is to test the hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$. We know that $Var(\hat{\beta}_j) = \sigma^2 C_{jj}$. Exploiting this fact, consider the test statistic

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Above t_0 follows a t-distribution with $n - k - 1$ degrees of freedom.

1.4.3 Extra Sum of Squares of Method

We just saw the test statistic which predicts whether the contribution of the variable x_i *individually* is significant or not. But many times we wish to know the significance of a variable x_i given that the model already has variables $x_1, x_2 \cdots, x_n$. For example, if we are predicting the height of a baby at some future time "t", then the variables x_1 = average height of relatives of the baby and x_2 = height of father are important variables *individually*. On the other hand, it is clear that the it is unnecessary to add both variables to the model i.e. x_1 is not significant given x_2 is already present. We wish to address this issue in the current section.

Extra Sum of Squares

Extra Sum of Squares method is used to determine if some subset of $r < k$ regressors contribute significantly to the model. Without loss of generality let them be the last r out of the k regressors we have. Then, the hypothesis we are testing is

$$\bar{\beta} = (\bar{\beta_1}, \bar{\beta_2})', \text{ then } H_0 : \bar{\beta_2} = 0$$

Here $\bar{\beta_2}$ is the vector containing subset of coefficients which we test. Now,

$$y_{n\times 1} = [X_1 X_2][\bar{\beta_1}, \bar{\beta_2}]' + \bar{\epsilon} = X_1\bar{\beta_1} + X_2\bar{\beta_2} + \bar{\epsilon} \text{ (this is called full model)}$$

Note that $X = [X_1 X_2]$ is the partitioned matrix with X_2 containing the values associated with the variables corresponding to coefficients in $\bar{\beta_2}$.

For the full model $\hat{\beta} = (X'X)^{-1}X'y$, $SSR = \hat{\beta}'X'y - \frac{(\sum_{i=1}^n y_i)^2}{n}$ and $MSRes = y'y - \hat{\beta}'X'y$.

When H_0 is true, then we get the reduced model $y = X_1\bar{\beta_1} + \bar{\epsilon}$. Then $\hat{\beta_1} = (X_1'X_1)^{-1}X_1'y$ and $SSR(\beta_1) = \hat{\beta_1}'X_1'y - \frac{(\sum_{i=1}^n y_i)^2}{n}$

Then the regression sum of squares due to β_2 given β_1 is given by $SSR(\beta_2|\beta_1) = SSR(\beta) - SSR(\beta_1)$. This quantity $SSR(\beta_2|\beta_1)$ is called the **extra sum of squares due to** $\bar{\beta_2}$ which gives the increase in regression sum of squares that results from adding regressors $x_{k-r+1} \cdots x_k$ to a model that already contains $x_1 \cdots x_{k-r}$

The partial F-test

First of all, $SSR(\beta_2|\beta_1) = y'[X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1']y$. Now 1.1 can be used to show that $\frac{SSR(\beta_2|\beta_1)}{\sigma^2} \sim \chi_r^2$ distribution. The using $SSRes(\bar{\beta}) = y'[I - X(X'X)^{-1}X']y$ and 1.3, we can conclude that $SSR(\beta_2|\beta_1)$ and $SSRes(\bar{\beta})$ are independent. Also, we know that

$$\frac{SSRes(\beta)}{\sigma^2} \sim \chi_{n-p}^2$$

Using the definition of F-statistic we can conclude that

$$\frac{SSR(\beta_2|\beta_1)}{r} \times \frac{1}{MSR(\bar{\beta})} \sim F_{r,n-p}$$

Chapter 2

Model Adequacy Checking

The following sources have been used to make this chapter [15], [pen] , [8] , [17] , [Clarke] , [4] , [Winner] , [Shalab] , [19], [Findsen and Troisi] , [14] and [2]

There were a set of assumptions that we made for our model in the previous chapter. The assumptions are

1. There exists a linear relationship between the regressors and response variable.
2. Errors are normal iid random variables.

Violations of these assumptions may yield to an unstable model and hence it is necessary to check if the above assumptions hold for our model. Before going through the specific tools which help us identify the violations from required assumptions, we will see about scaling residuals which help us predict potential outliers and influential data points.

2.1 Influential points, outliers, leverage and Scaling Residuals

The definitions given below may not the most precise definitions but they definitely help one to get an intuition about what the terms are.

An **outlier** is a data point whose response y does not follow the general trend of the rest of the data. On the other hand data point is said to have a **high leverage** if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values. Finally a data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.

It is important to understand that outliers and high leverage data points *have the potential* to be influential, but we generally have to investigate further to determine whether or not they are actually influential. To understand these terms better, let us see each term in slightly more detail.

2.1.1 Leverage

First let us understand leverage. Our first step will be to quantify the amount of leverage associated with a given point. Now we know that:
$\hat{Y} = X\hat{\beta} \implies \hat{Y}_i = \sum_{i=1}^{n} h_{ij} Y_j$ where $H = (h_{ij})$ is the hat matrix such that $\hat{Y} = HY$.
Observe that the fitted $\hat{Y}_i$ at i^{th} row changes at the rate $h_i j$ with respect to change in j^{th} observed value Y_j. So the j^{th} column of H, call it $H_j = [h_{1j}, h_{2j}, \cdots, h_{nj}]$ is the vector having all rates of changes with respect to j_{th} observations. Hence H_j is called the **leverage vector**.

Now, the length of vector H_j gives a measure of overall rate at which j^{th} observations change the value $\hat{Y}$. Since H is symmetric and idempotent, we have the following equality

$$\|H_j\|^2 = h_{1j}^2 + \cdots h_{nj}^2 = h_{jj}$$

Hence the leverage associated with the j^{th} observation is simply h_{ii} Some properties about leverage points are

1. $\|H_j\|^2 = h_{jj}$
2. $0 \leq h_{jj} \leq 1$
3. trace(H) = sum of all leverages.

Note that it is necessary to identify leverage points because they help us identify influential points. Also, it can be shown that $\hat{\epsilon} = (I - H)\epsilon$. Hence from the variance-covariance matrix of $\hat{\epsilon}$ we can say that $Var(\hat{\epsilon_i}) = \sigma^2(1 - h_{ii})$. Hence a point with high leverage produces a corresponding estimated residual with small variance.

Now we try to detect leverage points for a data set using R. The data we use is stored as "LifeCycleSavings" in R data sets.

```
> data("LifeCycleSavings")
> print(head(LifeCycleSavings))

             sr pop15 pop75     dpi ddpi
Australia 11.43 29.35  2.87 2329.68 2.87
Austria   12.07 23.32  4.41 1507.99 3.93
Belgium   13.17 23.80  4.43 2108.47 3.82
Bolivia    5.75 41.89  1.67  189.13 0.22
Brazil    12.88 42.19  0.83  728.47 4.56
Canada     8.79 31.72  2.85 2982.88 2.43

# Fit the data into the model according to " life-cycle savings hypothesis"
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)

> x <- model.matrix(g)
> lev <- hat(x) ## h_ii values

> plot(lev,ylab="Leverages",main="Index plot of Leverages") ## check Figure 2.1
> abline(h=2*5/50)  ## 2p/n line is rule of thumb

> countries <- row.names(LifeCycleSavings)
>names(lev) <- countries
>names(lev)

>lev[lev > 0.2] ##using rule of thumb to detect outlier
   Ireland         Japan       United States    Libya
    0.2122363     0.2233099     0.3336880      0.5314568
```

Hence the observations associated with Ireland, Japan, US and Libya seem to be having some potential to be influential points.

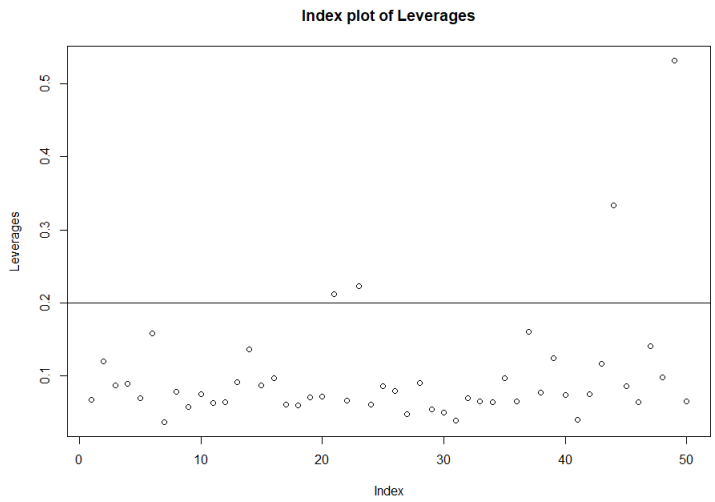


Figure 2.1: Plotting leverage values

2.1.2 Outliers

Outliers are those which simply do not follow the "trend" of the rest of the data. In order to mathematically identify an outlier, we need a test-statistic, which we get based on the R-student that we discussed earlier.

Assume that the model that we have initially fit the model with is $\bar{y} = X\bar{\beta} + \bar{\epsilon}$ when the actual model is assumed to be $\bar{y} = X\bar{\beta} + \bar{\delta} + \bar{\epsilon}$ where $\bar{\delta} = [0, 0 \cdots \delta_u \cdots 0]$ is a $n \times 1$ vector with zeros at all places except at the u^{th} position. The second model is called the Mean Shift Outlier Model. Both the actual and fitted models follow the normality and constant variance assumption for their errors. We wish to test the null hypothesis $H_0 : \delta_u = 0$ against $H_1 : \delta_u \neq 0$.

Remark 2.1. It is important to note that this test assumes that we are specifically interested in the u^{th} observation i.e. we have a *priori information* that the u^{th} observation might be an outlier.

Our first step is to estimate δ_u. An outline of how the statistic is derived is given below

$$\bar{e} = [I - H]\bar{y} \implies E(\bar{e}) = [I - H][X\bar{\beta} + \bar{\delta}] = (I - H)\bar{\delta} \implies E(e_u) = (1 - h_{uu})\delta_u$$

So a possible estimate for δ_u is $\hat{\delta_u} = \frac{e_u}{1-h_{uu}}$ which is nothing but the u^{th} PRESS residual $e_{(i)}$
$\bar{e} = [I - H]\bar{y}$ can be used to show that $Var(\hat{\delta_u}) = \frac{\sigma^2}{1-h_{uu}}$
All the information collected so far is used to show that under null hypothesis

$$\frac{e_u}{\sigma\sqrt{1 - h_{uu}}} \sim N(0, 1) \text{ (R-Student)}$$

We use $S_{(u)}$ to estimate σ which is the estimate of *sigma* with u^{th} observation deleted. This has been preferred over the use of $MSRes$ because e and $S_{(u)}$ are independent of each other.

Conclusion: If the u^{th} observation is not an outlier, then the R-student follows a standard normal distribution.

Now we put this test into use in R. We again use our "LifeCycleSavings" data set to detect possible outliers in the data set.

```
> jack <- rstudent(g) ## computes all R-student values
> plot(rstudent(g))
> jack[abs(jack)==max(abs(jack))]
  Zambia
2.853558
```

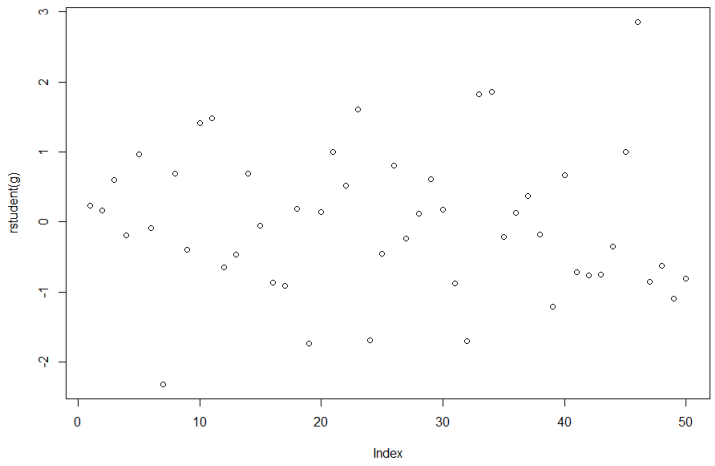


Figure 2.2: R-student values for countries indexed by numbers 1-50

At $\alpha = 0.05$ level of significance, the critical value for normal random variable is $z_{\alpha/2} = 1.96$. Since $2.853558 > 1.96$, we do not have enough evidence to accept the null hypothesis. Hence the observation corresponding to Zambia is an outlier.

Remark 2.2. Observe that :

- Even though our test showed evidence for Zambia being an outlier, it is to be noted that observations corresponding to Zambia were *not* one of the high leverage points.
- Similarly printing all the R-student values, we can observe that Libya has R-student value to be $-1.08930326 < 1.96$. Hence at 0.05 level of significance, we do not have enough evidence to believe that it is an outlier. Hence it is not an outlier but it is a high leverage point.

2.1.3 Influential Points

An influential point is one whose removal from the dataset would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have at least one of those two properties. We use an example to observe the behaviour of an influential point.

The example we are using is of a dataset with multiple outliers. Data available is that of log(surface temperature) and the log(light intensity) of 47 stars in the star cluster CYG OB1.

```
> library("HSAUR3")
> data("CYGOB1")

#temperature and the log of the light intensity of 47 stars in the star
cluster CYG OB1
> print(head(CYGOB1))
      logst logli
[1,]  4.37  5.23
[2,]  4.56  5.74
[3,]  4.26  4.93
[4,]  4.56  5.74
[5,]  4.30  5.19
[6,]  4.46  5.46
```

```
> plot(CYGOB1$logst, CYGOB1$logli,xlab="log(Surface Temperature)",
ylab="log(Light Intensity)")
#observe influential points in the plot

> model1 <- lm(CYGOB1$logli ~ CYGOB1$logst , data = CYGOB1)
> abline(model1)
> model2 <- lm(CYGOB1$logli ~ CYGOB1$logst , data = CYGOB1 ,
subset = (CYGOB1$logst > 3.6) )
> abline(model2 )   ## check Fig 2.3
```

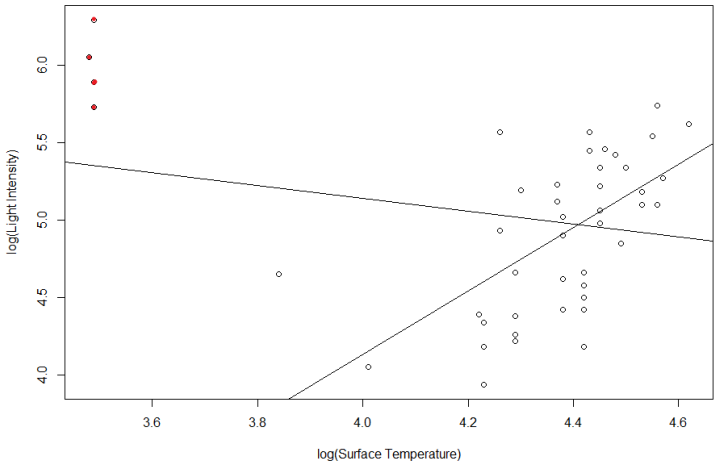


Figure 2.3: Influential points in red

First we draw a line considering all the data points (including the red ones) and the second line is drawn excluding the four red points. It is clear by comparing the two lines that why the red points are called "influential".

Now how do we quantify influence. One method is to use the cook's D-statistic. It is given by

$$D_i = \frac{[\hat{\beta} - \hat{\beta_{(i)}}]'(X'X)[\hat{\beta} - \hat{\beta_{(i)}}]}{n\hat{\sigma}^2}$$

Higher values of this statistic give evidence that the point is influential. Let us go back to our LifeCycleSavings dataset.

```
> cook <- cooks.distance(g)
> plot(cook,ylab="Cook's distances")
```

Having obtained the three countries which probably might be influential, we will try to check which one actually is influential. Consider Japan for instance. Let us consider two models, one with Japan included and another excluding Japan.

```
> summary(g) # model with all points including japan

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
```

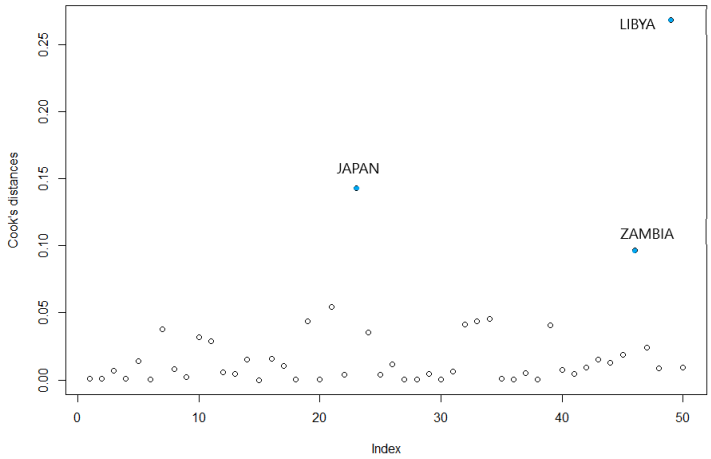


Figure 2.4: Points with high cook's distance in red

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904

> gj <- lm( sr ~ pop15 + pop75 + dpi + ddpi ,data = LifeCycleSavings ,
subset=(countries != "Japan"))
> summary(gj)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings,
    subset = (countries != "Japan"))

Residuals:
   Min     1Q Median     3Q    Max
-7.997 -2.592 -0.115  2.032 10.157

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.9401714  7.7839968   3.076  0.00361 **
pop15       -0.3679015  0.1536296  -2.395  0.02096 *
pop75       -0.9736743  1.1554502  -0.843  0.40397
```

```
dpi          -0.0004706  0.0009191  -0.512  0.61116
ddpi          0.3347486  0.1984457   1.687  0.09871 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.738 on 44 degrees of freedom
Multiple R-squared:  0.277,Adjusted R-squared:  0.2113
F-statistic: 4.214 on 4 and 44 DF,  p-value: 0.005649
```

Observations: From the above two summaries, we see that removing Japan from our data set has made the variable "ddpi" insignificant while it was initially significant in the full model.
Conclusion : Hence Japan does seem to be influencing our model. It seems to be an influential point. Similarly we can check the other countries.

2.1.4 Scaling Residuals

Residuals are scaled in order to detect outliers or extreme values. Raw residuals can have any range of values and hence in such a situation defining "extreme" becomes difficult. Some of the ways of scaling residuals are :

Standardized Residuals

The standardized residual is given as below which follows an approximate normal distribution.

$$d_i = \frac{e_i}{\sqrt{MSRes}} \approx N(0,1)$$

Here we have estimated variance of e_i using $MSRes$. Note that large values of residuals indicate potential outliers.

Studentized Residuals

Instead of using $MSRes$ to estimate σ^2, here we attempt to find the actual variance of e_i. It can be shown that:

$$e = (I-H)y = (I-H)(X\beta+\epsilon) = (I-H)\epsilon \implies Var(e) = \sigma^2(I-H) \implies Var(e_i) = \sigma^2 h_{ii} \tag{2.1}$$

Two points worth noting from Equation 2.1 are :

- Error vector e is associated with the residuals vector ϵ via the linear transformation $I-H$.
- Even though errors ϵ_i are assumed to be uncorrelated and independent, the residuals e_i are correlated and hence dependent.

When σ is unknown, we estimate variance σ^2 using $MSRes$. Hence our unbiased estimate for $Var(e_i)$ is $\sqrt{MSRes(1-h_{ii})}$. Hence our Studentized residuals are:

$$r_i = \frac{e_i}{\sqrt{MSRes(1-h_{ii})}} \nsim \text{ t-distribution}$$

Note that r_i *does not* follow a t-distribution because $MSRes$ and e_i are not independent of each other. But it is still called "studentized" because we have have scaled the residual using estimated standard deviation.

Remark 2.3. Studentization can only correct for the natural non-constant variance in residuals when the errors have constant variance. If there is some underlying heteroscedascity in the errors, studentization cannot correct for it.

PRESS Residuals

We know that a leverage point "drags" the prediction line to itself. This implies that e_i will be small and hence we won't be able to detect the outlier. $PRESS$ residuals have been developed to tackle this problem.

$$e_{(i)} = y_i - \hat{y_{(i)}} = \frac{e_i}{1 - h_{ii}} = \text{ PRESS residual} \tag{2.2}$$

We will later show that $e_{(i)} = \frac{e_i}{1-h_{ii}}$ and hence the standardized $PRESS$ residuals can be defined as:

$$\frac{e_{(i)}}{\sqrt{Var(e_{(i)})}} = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}}$$

If we estimate σ^2 by $MSRes$, then we get the studentized residuals.This procedure is called internal scaling of residuals. Another approach to is to estimated σ^2 by deleting the i^{th} observation. This estimate is given by

$$S^2_{(i)} = \frac{(n-p)MSRes - \frac{e_i^2}{1-h_{ii}}}{n-p-1}. \tag{2.3}$$

The proofs of Equations 2.2 and 2.3 are lengthy but not hard. One can refer to pp 591-594 of the text [15] for a detailed proof.
When $S^2_{(i)}$ is used to estimate variance, then it produces an externally studentized residual, also called as the **R-student**.

2.2 PRESS statistic

The quantity $e_{(i)} = y_i - \hat{y_{(i)}}$ is called as the **PRESS Residual**. Now the $PRESS$ statistic is defined as

$$PRESS = \sum_{i=1}^{n} [y_i - \hat{y_{(i)}}]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}}\right)^2$$

$PRESS$ statistic is a measure of how well a regression model will perform in predicting new data. Observe that if a data point has high leverage then the $PRESS$ statistic value gets large. So large $PRESS$ values indicate the presence of high leverage points which have the potential to be influential. Hence models with lower $PRESS$ values are desired because they predict future observations more accurately.

$PRESS$ can be used to detect an R^2 like statistic which measures how well the model fits the data. The definition of R^2 we know is

$$R^2 = 1 - \frac{SSRes}{SST}$$

If y_i is an influential point, then it it drags the prediction line towards itself. Hence $y_i - \hat{y}_i$ will be small. This means that the effect of the influential point is not visible it the $PRESS$ statistic. To avoid this problem, we define adjusted R^2 as :

$$R^2_{adj} = 1 - \frac{PRESS}{SST}$$

2.3 Constant Variance Assumption

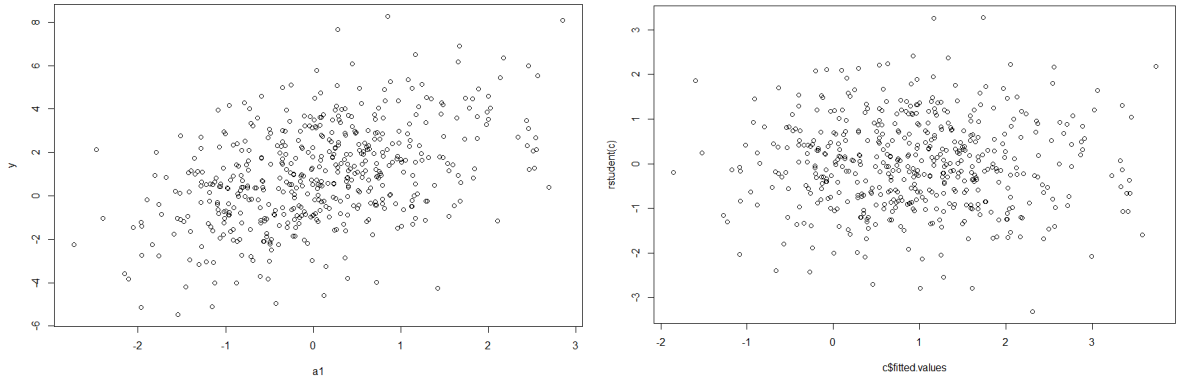
We earlier saw that $y_i - \hat{y}_i$ is called as the i^{th} residual. We use the plot between fitted values and various kinds of residuals to get a brief idea about whether or not the model follows constant variance.

Fitted values vs. Raw residuals

Here we plot "Fitted values" on x-axis and "raw-residuals" on the y-axis and depending on the pattern observed, we get a brief idea about the error variance.

First we consider a case where the constant variance assumption is met.

```
a1 <- rnorm(500, 0 , 1) #std normal
eps2 <- rnorm(500, 0,2)
y = 1 + a1 + eps2
c <- lm(y~a1)
plot(c$fitted.values, rstudent(c))
```



(a) a1 versus y plot (b) fitted values versus R-student

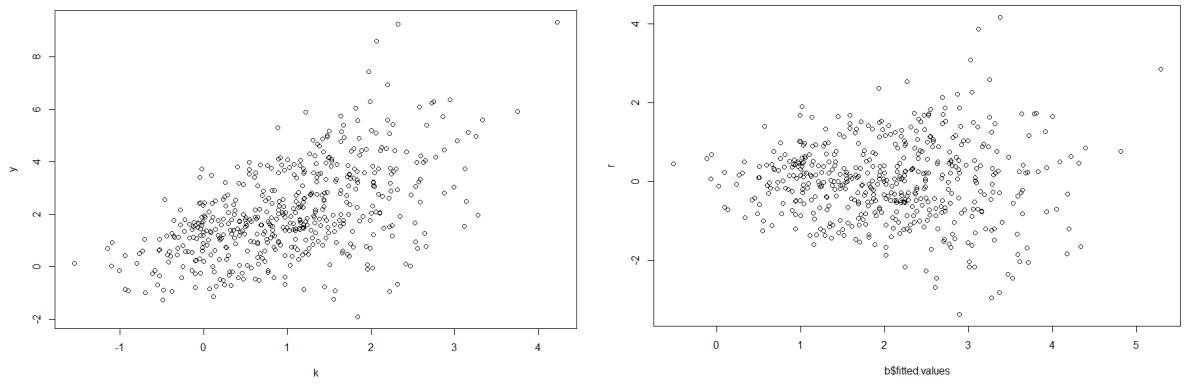
Figure 2.5: Residual plot for model with constant error variance

Now we consider the case of non-constant variance

```
seed(0)
k = rnorm(500,1,1)
b0 = 1 # intercept chosen at your choice
b1 = 1 # coeff chosen at your choice
h = function(k) 1+ 0.4*k
## h performs heteroscedasticity function
eps = rnorm(500,0,h(k))
y = b0 + b1*k + eps
plot(k,y)
b <- lm(y~k)
r <- rstudent(lm(y~k))
plot(b$fitted.values, r )
```

Now, let us go back to "LifeCycleSavings" dataset.

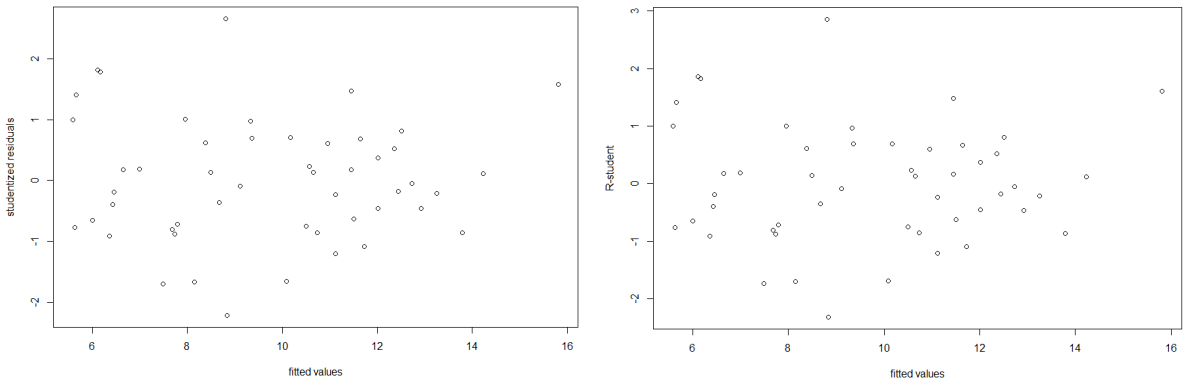
```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings) ## to recall
> gs <- summary(g)

# better to draw conclusions after the residuals have been normalized
> s <- gs$sigma
> lev <- hat(x)
> stud <- g$residuals/(s*sqrt(1-lev)) #studentised residulas
> plot(g$fitted.values , stud, xlab = "fitted values" ,
ylab = "studentized residuals")
```



(a) k versus y plot (b) fitted values versus R-student

Figure 2.6: Residual plot for model with non-constant error variance

```
#jacknife residuals or R-student
> jack <- rstudent(g)
> plot(g$fitted.values , jack , xlab = "fitted values" , ylab = "R-student")
```



(a) fitted values vs. studentized residuals (b) fitted values vs. R-student

Figure 2.7: Residual plot for LifeCycleSavings dataset

Conclusion: Observing the residual plots in Figure 2.7, we can say that the model we have assumed appears to be satisfying the constant variance assumption.

2.3.1 Variance stabilizing transformations

In our "LifeCycleSavings" data set, we did not observe the problem of non-constant variance. But what if a model exhibits the problem? How do we deal with it? This question is answered by applying a transformation to our response variable, so that the transformed model satisfies the constant variance assumption.

One of the common reasons for the violation of constant variance is for the response variable to follow a distribution in which variance is a function of mean i.e. when $\sigma^2 = \omega(\mu)$.
AIM: We wish to find a function f such that $Var(f(Y))$ is roughly constant i.e. we "transform" the response variable.

$$f(Y) \approx f(\mu) + (Y - \mu)f'(\mu) \Rightarrow [f(Y) - f(\mu)]^2 \approx (Y - \mu)^2[f'(\mu)]^2$$

$$\text{Hence } \boxed{V(f(Y)) \approx V(Y) \times [f'(\mu)]^2} \tag{2.4}$$

Using (2.4), we find a transformation f such that $Var(f(Y))$ is constant.

Example 2.1. For poisson distribution $Var(Y) = \omega(\mu) = \mu$.

$$[f'(\mu)]^2 = \frac{1}{\mu} \Rightarrow f(\mu) = 2\sqrt{\mu}$$

△

2.4 Normality assumption

Now, we move to the normality assumption and how to detect its violation. This assumption says that all the errors are iid normal random variables.

2.4.1 Q-Q plot

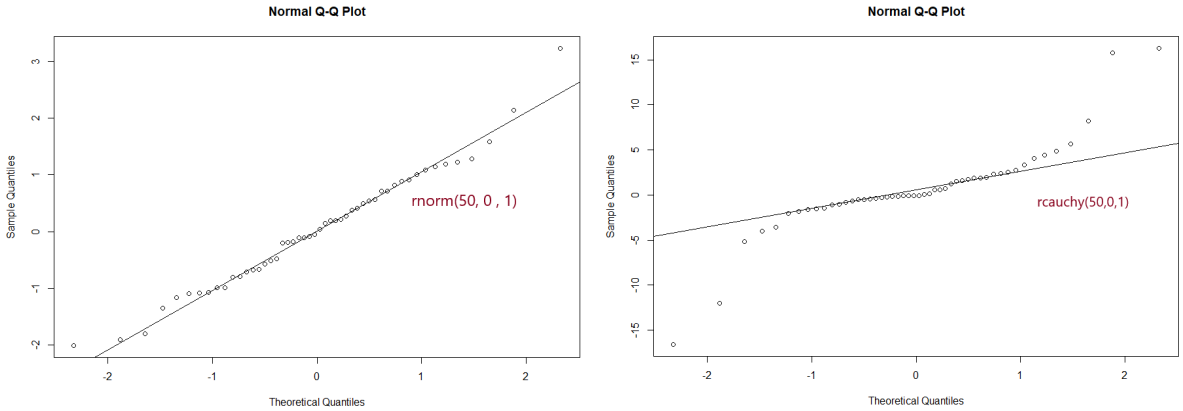
Q-Q plot is a graphical tool that is used to assess normality. It plots the sample quantiles (vertical axis) against the theoretical quantile (horizontal axis) where the original data point x_i value is called sample quantile and the expected z-score for the data point x_i is called theoretical quantile. If the data comes from a normal distribution, then the theoretical and sample quantiles match, hence giving an approximately straight line Q-Q plot. Otherwise, the plot is not straight line.

A set of different R-codes are given below to obtain Q-Q plot for points coming from different kinds of distributions. The pictures of the corresponding Q-Q plots is also attached after the codes.

```
a1 <- rnorm(50, 0 , 1) #std normal
qqnorm(a1)
qqline(a1)

c <- rcauchy(50,0,1) # heavy tailed distribution
qqnorm(c)
qqline(c)

skewR<- rchisq(50, df = 3) # skewed to right
qqnorm(skewR)
#qqplot(qnorm(ppoints(30)) , qchisq(ppoints(30),df=3)) # skewed to the right

skewL <- rbeta(50,2,0.5,ncp=2) # skewed to left
qqnorm(skewL)

bi <- rbeta(50, 0.5 ,0.5) # bimodal distribution
qqnorm(bi) ##observe the gap near zero
```

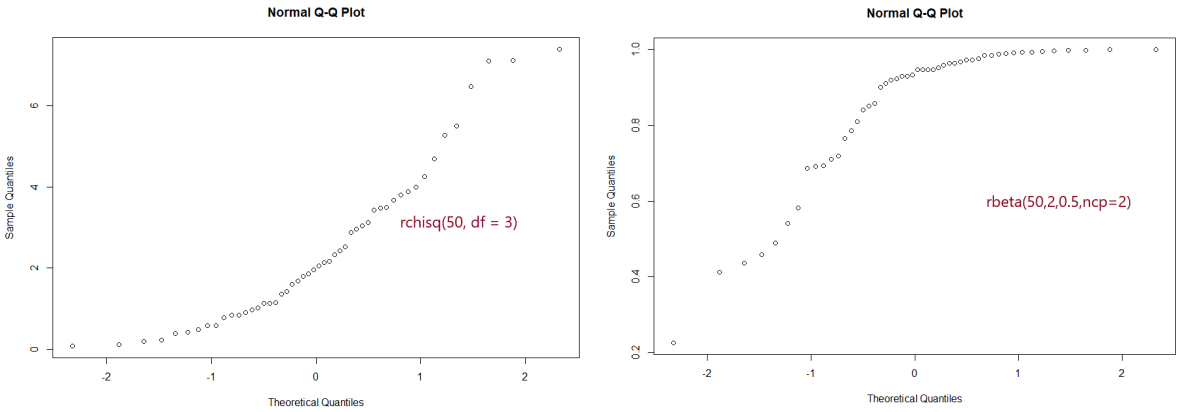
Now we consider our "LifeCycleSavings" data set and check if its Q-Q plot exhibits any violation from normality. The Q-Q plot for this datais given in 2.11

```
# g is our linear model, defined earlier
qqnorm(rstudent(g))
abline(0,1) ## a,b = intercept and slope values
```



(a) Normal distribution (b) Heavy-tailed distribution

Figure 2.8



(a) Skewed to the right (b) Skewed to the left

Figure 2.9

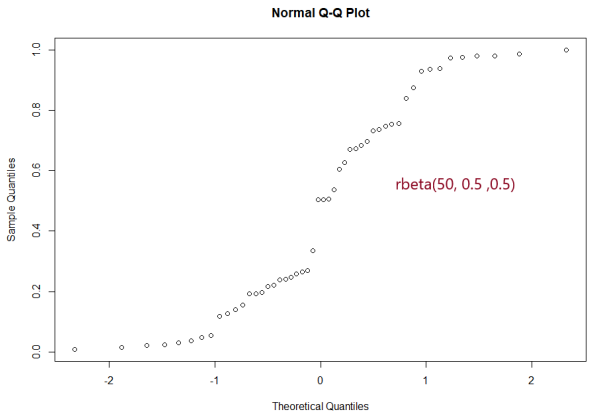


Figure 2.10: Bi-modal Distribution

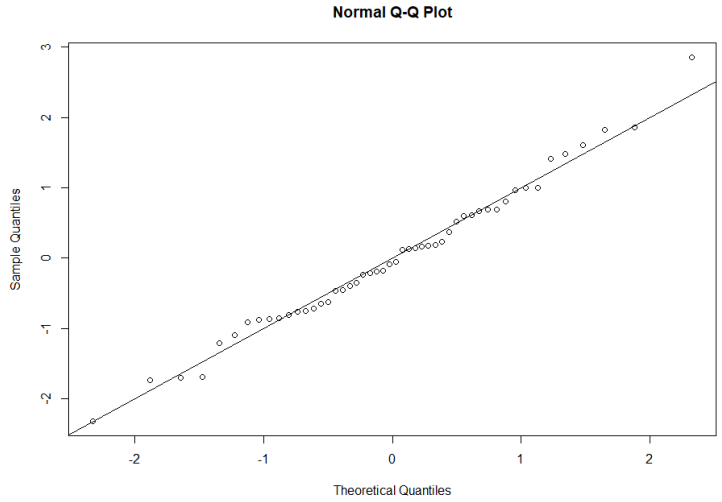


Figure 2.11: Q-Q plot for LifeCycleSavings dataset

2.4.2 Tests for Normality

Kolmogorov-Smirnov Test

If $X_1, X_2 \cdots X_n$ are assumed to come from a continuous distribution P. Then we want to test the null hypothesis H_0 : The samples come from P against H_1 : they do not come from P.

If F is the cdf of X under H_0, then the empirical distribution function F_{obs} is

$$F_{obs} = \frac{I(X < x)}{\text{total observations}}$$

Let F_{exp} be the cdf associated with the null hypothesis. Then, then Kolmogorov-Smirnov statistic is given by

$$\boxed{D_n = max\{\mid F_{exp}(x) - F_{obs}(x) \mid\}}$$

Procedure:

- The first step is to order the data. Let the ordered data be $x_{(1)}, \cdots x_{(n)}$ so that $F_{obs}(x_{(i)}) = i/n$.
- Find $F_{exp}(x_{(i)})$ for each i. Now tabulate the values of $\mid F_{exp}(x) - F_{obs}(x) \mid$ for x-value. Maximum of all these values is given by D_n.
- Critical value for $\alpha = 0.05$ is given by $D_{c,0.05} = \frac{1.36}{\sqrt{n}}$.

Hence we reject the null hypothesis if $D_n > D_{c,0.05}$. Now, we go back to LifeCycleSavings dataset

```
> ks.test(as.numeric(rstudent(g)), pnorm)
One-sample Kolmogorov-Smirnov test

data:  as.numeric(rstudent(g))
D = 0.067991, p-value = 0.9628
alternative hypothesis: two-sided
```

From the p-value that we have obtained, we can say that the normality assumption is being followed.

Anderson-Darling Test

This is also a test to check if a sample has come from a specific distribution. This is a modification of the KS test. In KS test, critical value does not depend on the specific distribution being tested but on the other hand AD test makes use of the specific distribution compute critical values.

The test statistic used here is

$$A^2 = -N - \sum_{i=1}^{N} \frac{2i-1}{N} \left(ln(F(Y_i) + ln(1 - F(Y_{N+1-i}))\right)$$

Where Y_i is the ordered data, F is the cdf of the specified distribution.

Shiapiro-Wilk Test

The Shapiro-Wilk test, proposed in 1965, calculates a W statistic that tests whether a random sample, $X_1, X_2, \cdots, X_n$ comes from a normal distribution. The W statistic is given by:

$$W = \frac{\sum_{i=1}^{n}(a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where $X_{(i)}$ are the ordered sample values , $(a_1, a_2, \cdots a_n) = \frac{m'V^{-1}}{c}$ for $c = \parallel V^{-1}m \parallel$. Here V is the varance-covariance matrix of the order statistics and $m = (E(X_{(1)}), \cdots, E(X_{(1)}))^T$.

Cramer Von Mises Test

Assume that μ and $sigma$ are unknown but are estimated from the data, then the cramer von mises test statistic is

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F(X_i) - \frac{21-1}{2n} \right)^2$$

2.5 Box-Cox Transformation

If there is evidence of non-normality, then the standard remedy is to transform the response using box-cox transformation. All response variables in box-cox method need to be positive. This method assumes that there exists transformation parameter λ such that $Y_I^{(i)} = g(Y_i, \lambda) = x_i'\beta + \epsilon$ where

$$g(Y_i, \lambda) = \begin{cases} \frac{Y_i^{\lambda-1}}{\lambda} & \text{if} \lambda \neq 0 \\ logY_i & \text{if} \lambda = 0 \end{cases}$$

Assuming that the transformed response $f(Y)$ follows multivariate normal distribution, we try to maximize the likelihood function of response variable w.r.t λ. The details are given in montgomery book

We earlier saw that the LifeCycleSavings data appeared to follow the normality assumption. Hence there is no need to apply the box-cox transformation to this data. Given below is an a data set where we have to apply the box-cox transformation.

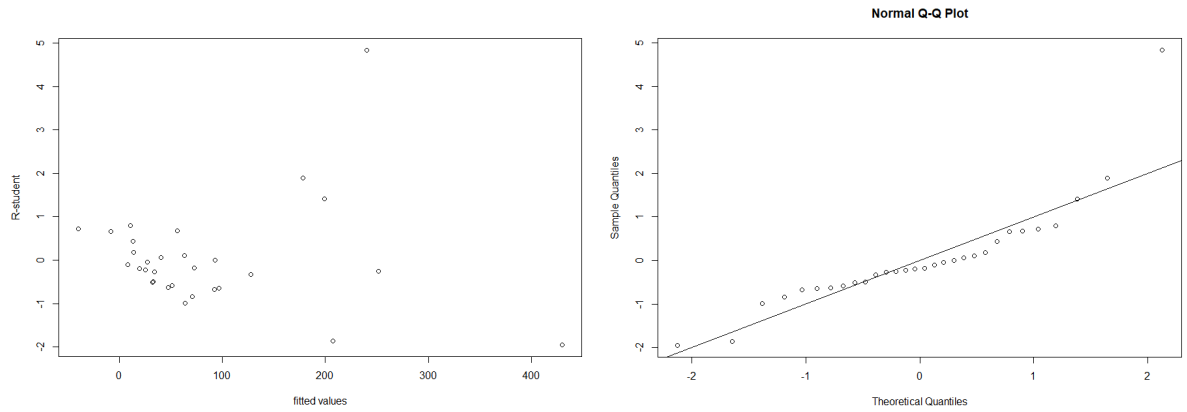
We use the "gala" data from the R "faraway package" which includes data for seven variables for each of the thirty Galapagos islands. The relationship between the number of plant species in the place and several geographic variables is given in the data. The original dataset contained several missing values which have been filled for our convenience.

```
> library(faraway)
> data(gala) # relationship between number of plant species and other geographic
# variables of interest

> head(gala)
            Species Endemics  Area Elevation Nearest Scruz Adjacent
Baltra           58       23 25.09       346     0.6   0.6     1.84
Bartolome        31       21  1.24       109     0.6  26.3   572.33
Caldwell          3        3  0.21       114     2.8  58.7     0.78
Champion         25        9  0.10        46     1.9  47.4     0.18
Coamano           2        1  0.05        77     1.9   1.9   903.82
Daphne.Major     18       11  0.34       119     8.0   8.0     1.84

> dim(gala) # 30 islands and 7 variables
[1] 30  7

> gfit <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, data=gala)

> summary(gfit) # adjusted R^2 = 0.71
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07

# we can remove the insignificant variables

> gfit2 <- lm(Species ~ Elevation + Adjacent,
            data=gala)

> summary(gfit2)
Call:
lm(formula = Species ~ Elevation + Adjacent, data = gala)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-103.41  -34.33  -11.43   22.57  203.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.43287   15.02469   0.095 0.924727
Elevation    0.27657    0.03176   8.707 2.53e-09 ***
Adjacent    -0.06889    0.01549  -4.447 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.86 on 27 degrees of freedom
Multiple R-squared:  0.7376,Adjusted R-squared:  0.7181
F-statistic: 37.94 on 2 and 27 DF,  p-value: 1.434e-08

> summary(gfit2) # adjusted R^2 = 0.71 , SE = 60.86
```

After reducing the data to some extent, we now move to checking for the variance and normality assumptions. The Residual plot and Q-Q plot corresponding to the data is given in 2.12

```
### checking for non constant variance
> jack <- rstudent(gfit2)
> plot(gfit2$fitted.values , jack , xlab = "fitted values" , ylab = "R-student")
# we suspect non constant variance

> qqnorm(rstudent(gfit2)) # looks like heavy tail distribution
> abline(0,1) # slightly skewed distribution of error
## so transformation is required
```



(a) Residual plot for gala data

(b) Q-Q plot for gala data

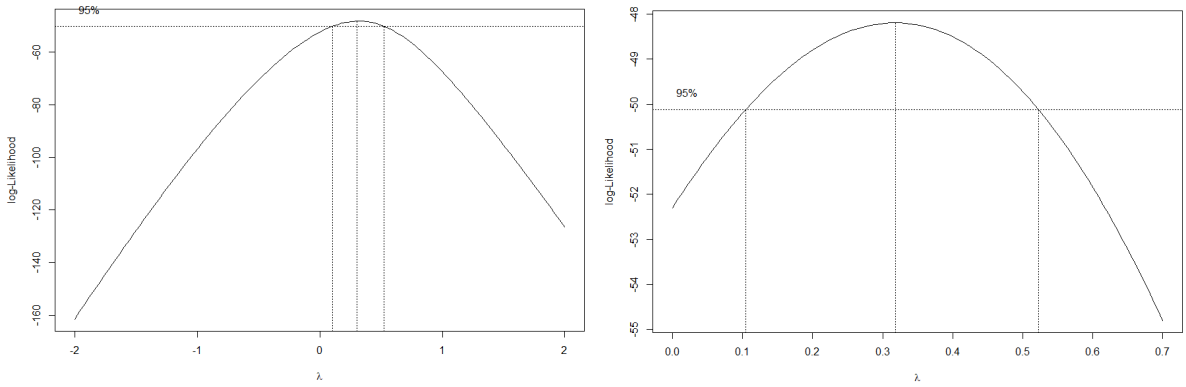
Figure 2.12

The “MASS” package needs to be installed on R in order to apply the box-transformation. The “boxcox()” function in R plots the log-liklihood function as a function of λ. From the plot, we can identify the approximate value of λ that maximizes the likelihood function.

```
> library(MASS)
> boxcox(gfit2, plotit = T)
> boxcox(gfit2,lambda=seq(0.0,0.7,by=0.05),plotit=T)
# approx lambda which maximizes log likelihood is 0.3
```

From Figure 2.13, we can see that $\lambda \approx 0.3$. This suggests that we have to choose a cube root transformation to our data. Once the cube root transformation is applied, we again have to check for the model assumptions. This has been done in the R-code given below.

```
# we can try cube root transformation
> gfit3 <- lm((Species)^(1/3) ~ Elevation + Adjacent,
            data=gala)

> sumary(gfit3) # R squared = 0.69  , SE = almost 0 (reduced)

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  2.35243885  0.25428722  9.2511 7.340e-10
Elevation    0.00413487  0.00053757  7.6918 2.843e-08
Adjacent    -0.00088902  0.00026216 -3.3911  0.002159

n = 30, p = 3, Residual SE = 1.03002, R-Squared = 0.69

> jack3 <- rstudent(gfit3)
> plot(gfit3$fitted.values , jack3 , xlab = "fitted values" ,
ylab = "R-student")
# mostlty every point between -2, +2 and only two points outside this range
> qqnorm(rstudent(gfit3))
```

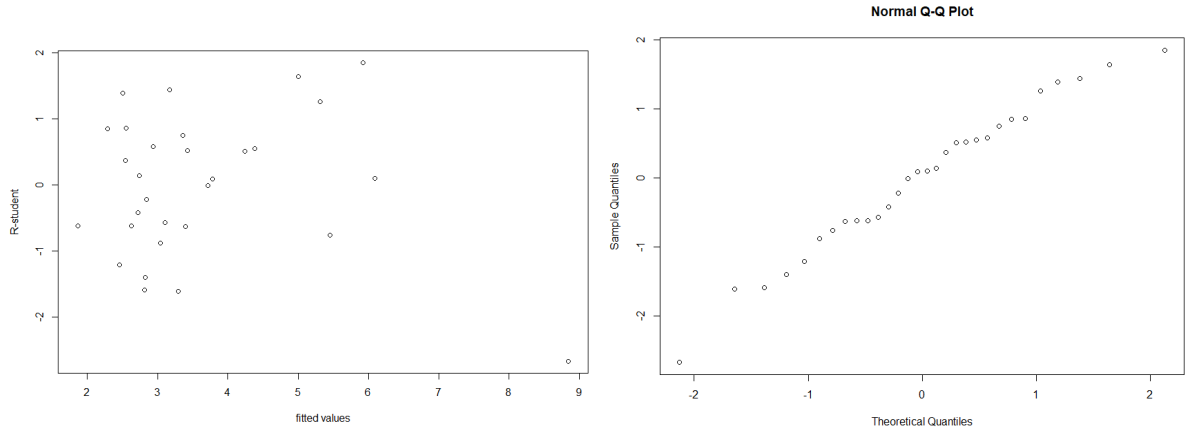


(a) Log-likelihood vs λ (b) Zoomed in likelihood plot

Figure 2.13

Observations:
Figure 2.14 shows the residual plot and Q-Q plot for the transformed model. It is clear that the residual plot appears more spread for the transformed model when compared to that of the model "gfit2". Also, the Q-Q plot appears to have become slightly more linear (earlier, the plot looked a little like heavy tailed distribution.) All these descriptive plots give us some evidence that the transformation did help resolve the normality and variance issue.

But we need to believe in numbers, the plots only give us a brief idea of what is happening. Consider the following table:



(a) Residual plot for transformed model

(b) Q-Q plot for the transformed model

Figure 2.14

Model	gfit2	gfit3
R^2 / adj R^2	0.71	0.69
Std. Error	60.86	1.03002

Table 2.1: Comparing the initial model "gfit2" and transformed model "gfit3"

Observe in Table 2.1 the huge decrease in the standard error for the new transformed model which still has a considerably good R^2 value. Let us now use Kolmogorov-Smirnov test to this new data to test for normality.

```
> ks.test(as.numeric(rstudent(gfit3)), pnorm)

One-sample Kolmogorov-Smirnov test

data:  as.numeric(rstudent(gfit3))
D = 0.093249, p-value = 0.935
alternative hypothesis: two-sided
```

From the p-value above, we can say that the normality assumption is being followed very well and hence this is a very good model to do regression analysis on.

Chapter 3

Multi-collinearity

The following sources have been used to make this chapter [15], [pen] and [10].

Suppose, we wish to fit the model : $Y = X\beta+\epsilon$. Then we know that the least square estimates are given by $(X'X)^{-1}X'y$. Now assume that two or more regressor variables are dependent on each other, then two or more columns of the X matrix will be linearly dependent. This means that neither X nor $X'X$ are invertible. Hence there do not exist least square estimates for this model. In this chapter we will first see about the effects of multi-collinearity. We will also learn about identifying and diagnosing multi-collinearity.

Note that if there is no linear relationship between the regressors i.e. the columns are linearly independent then the columns are said to be **orthogonal**. On the other hand, if the columns of X exhibit near linear dependencies, then the problem of **multicollinearity** is said to exist. Hence the problem of multi-collinearity is said to exist when two or more regressor variables are strongly correlated.

Dimensionless regression coefficients are said to be standardized regression coefficients. This is useful to use because the value of the coefficient will then be independent of the units of the regressor variable involved. These coefficients will be used often in this chapter.

3.1 Standardized regression coefficients

Unit normal scaling

In this case, we transform the predictor values x_{ij} and the response values y_i as follows:

$$x_{ij} \longrightarrow z_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j} \text{ where } s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}{n-1}$$
$$y_i \longrightarrow y_i* = \frac{y_i - \bar{y}}{s_y} \text{ where } s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

It can be shown that after transforming the variables, the new model we get with standardized coefficients have no constant(β_0) term.

3.1.1 Unit length scaling

In this case, we transform the predictor values x_{ij} and the response values y_i as follows:

$$x_{ij} \longrightarrow w_{ij} = \frac{x_{ij} - \bar{x_j}}{\sqrt{s_{jj}}} \text{ where } S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2$$
$$y_i \longrightarrow y_i^0 = \frac{y_i - \bar{y_i}}{\sqrt{SST}}$$

In this case as well, transformation eliminates the constant term from the model. There area few other properties of this type of scaling that makes it very convenient to use. One of the most useful properties is that if $W = (w_{ij})$ is the matrix of transformed regressor variable values, then:

$$W'W = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{1k} & x_{2k} & x_{3k} & \dots & 1 \end{bmatrix} \; W'y^0 = \begin{bmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{ky} \end{bmatrix} \tag{3.1}$$

In the above matrices, r_{jy} is the simple correlation between the regressor x_j and the response variable y. Also, r_{ij} is the correlation between regressors x_i and x_j. Another important point to note is that the regression coefficients that we compute in both cases (unit normal and unit length scaling) will turn out to be the same because we made sure our coefficients turn out to be dimensionless.

3.2 Inflation in variance

One of the effects of multicollinearity is high variance of the least square estimates. To see this, let us first consider the simple case where there are only two regressor variables x_1 and x_2. After scaling it to unit length as in 3.1.1, the model we get can be assumed to be $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$ (without loss of generality).

—

3.3 Detecting Multi-collinearity

Before developing tools for multicollinearity, I would like to mention some of the possible consequences of multi-collinearity which provide us hints about its existence.

- The t-test for each of the individual slopes are non-significant but the over all F-test for testing all of the slopes are simultaneously 0 significant.
- correlations among pairs of predictor variables are large. (this is not always reliable)

Now, we will see tools to detect multicollinearity in our data.

3.3.1 Correlation matrix and matrix scatter plot

The first step to do while we are trying to detect multicollinearity is to check if there are some very high pairwise correlations between the variables. For this we can either observe a correlation matrix or a matrix scatter plot as shown in the R-code below. The data has been take from the Penn State University STAT course website.

```
##MULTICOLLINEARITY IN DATA
# y = BP in mm Hg ; Wt in kg ; BSA in sq mt ; Dur in yrs ;
# Basal Pulse in beats/min ; Stress index
> data <- read.delim(choose.files(), header = T , sep = "\t")
> pairs(data, lower.panel = NULL)
> cor(data) ##symmetric matrix
                Pt         BP        Age     Weight         BSA       Dur     Pulse     Stress
Pt      1.00000000 0.03113499 0.04269354 0.02485650 -0.03128800 0.1762455 0.1122851 0.34315169
BP      0.03113499 1.00000000 0.65909298 0.95006765  0.86587887 0.2928336 0.7214132 0.16390139
Age     0.04269354 0.65909298 1.00000000 0.40734926  0.37845460 0.3437921 0.6187643 0.36822369
Weight  0.02485650 0.95006765 0.40734926 1.00000000  0.87530481 0.2006496 0.6593399 0.03435475
BSA    -0.03128800 0.86587887 0.37845460 0.87530481  1.00000000 0.1305400 0.4648188 0.01844634
Dur     0.17624551 0.29283363 0.34379206 0.20064959  0.13054001 1.0000000 0.4015144 0.31163982
Pulse   0.11228508 0.72141316 0.61876426 0.65933987  0.46481881 0.4015144 1.0000000 0.50631008
Stress  0.34315169 0.16390139 0.36822369 0.03435475  0.01844634 0.3116398 0.5063101 1.00000000
```

From the above code we can see that the "cor()" function in R gives a matrix containing all the correlations as its entries. Observe that this matrix has to be symmetric. Also a matrix plot is given for or data by the function "plot()". The function "plot" gives the scatter plot matrix of the given data. The picture of our data's scatter plot matrix is given in 3.1. The boxes highlighted with a red border exhibit high correlation while the boxes highlighted in green seem to exhibit moderately strong collinearity. Also the correlation matrix has some entries who values are above 0.7 indicating a significant amount of near linear dependencies among certain specific variables.

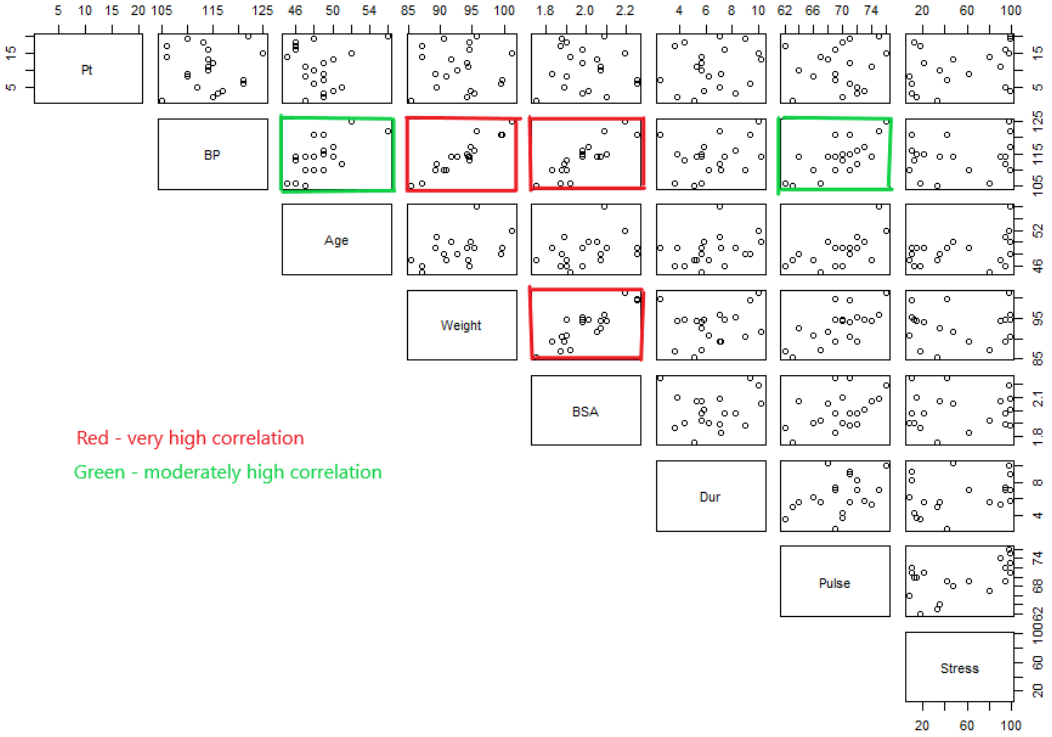


Figure 3.1: Scatter plot matrix for "blood pressure" dataset

3.3.2 Variation Inflation Factors(VIFs)

VIFs quantify how much of the variance is inflated due to the existence of multicollinearity. VIF exists for each of the predictors in a multiple regression model. VIF for the j^{th} predictor is given by VIF_j.

Rule of thumb: VIFs exceeding 4 warrant furthur investigation, VIFs exceeding 10 are signs of serious multicollinearity requiring correction. Using the unit normal scaling, we can get a new transformed model such that $X'X$ is the matrix containing correlations. Then we can show that $(X'X)^{-1}$ has diagonal elements equal to $C_{jj} = \frac{1}{1-R_{jj}^2}$.

Since $Var(\hat{\beta}_j) = C_{jj}\sigma^2$, we can say that variance of the parameters tend to infinity as R_j^2 tends to 1. Hence large variances and covariances for the least square estimators of the regression coefficient. So different samples taken at the same x levels could lead to widely different estimates of the model parameters. The followng is the R-code to get VIFs:

```
> a <- lm(data$BP ~ data$Weight + data$BSA)
> library(car)
> vif(a)
> data$Weight     data$BSA
```

```
4.276401    4.276401
```

From the VIF values above we can say that there is some evidence for multi-collinearity but further evidence is required to prove its existance.

3.3.3 Eigen System Analysis

Let $L_1 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$. Then we can show that

$$E(L_1^2) = \sum_{i=1}^{p} \sigma^2 Trace(X'X)^{-1}$$

We know that the trace of a matrix is the sum of eigen values of $(X'X)^{-1}$. When there is high multi-collinearity, $R_j \longrightarrow 1$ and hence $Tr(X'X)^{-1} \longrightarrow \infty$. So a matrix with high collinearity will also have the sum of eigen values of $(X'X)^{-1}$ also to be very high. This means that the sum of eigen values of $X'X$ is very small.

Condition Number: Condition number is defined as $K = \frac{\lambda_{max}}{\lambda_{min}}$. The rule of thumb is that if $K < 100$, then there is no problem with multi-collinearity. K between 100 and 1000 implies moderate to strong multi-collinearity. $K > 1000$ implies a severe multicollinearity.

The folloing is the R-code which helps do eigen system analysis on the blood pressure data set.

```
> eigen(cor(data))$values ## observe multicollinearity
[1] 3.919270634 1.640145177 0.878834608 0.706167046 0.474196726 0.301980238
    0.077250034 0.002155536

> kappa(cor(data) , exact = T) ## to get condition number
[1] 1818.234
```

Above condition number suggests the presence of severe multi-collinearity.

3.4 Problem to the solution: Ridge Regression

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. Here $\lambda > 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero. Hence the ridge regression problem is

$$\beta_{ridge} = argmin \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 \right\} \text{ subject to } \sum_{i=1}^{n} \beta_j^2 \leq s \tag{3.2}$$

By the method of Lagrange multipliers, this is equivalent to the following problem:

$$\beta_{ridge} = argmin \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 + \lambda \sum_{i=1}^{n} \beta_j^2 \right\} \tag{3.3}$$

There is a one-to-one correspondence between the parameters λ in (3.3) and s in (3.2). When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. A wildly large positive coefficient on one variable can be cancelled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients, this phenomenon is prevented from occurring.

The solution for (3.2) is

$$\beta_{ridge} = (X'X + \lambda I)^{-1}X'Y$$

The following is an R example for which we apply ridge regression. The data we used here is called the "meatspec" data. In this, we wish to estimate the fat content in a meat piece using the absorbance values that are recorded at 100 different wavelengths. There are 215 meat samples.

```
> library(faraway)
> data(meatspec , package = "faraway")
> mod1 <- lm(fat~ . , data = meatspec)

> summary(mod1)$r.squared  # observe R squared value
[1] 0.9951113

> kappa(mod1) ## very high condition number
[1] 1544436

> c = model.matrix(mod1)
> det(t(c)%*%c)
[1] 0
```

The determinant of the model matrix and the extremely high condition number for the matrix indicates presence of high multi-collinearity. So our aim would now be to apply ridge regression.

```
> library(MASS)
> require(MASS)
> lambda_seq = seq(0,5*10^(-8),len=21)
> ridgeregg <- lm.ridge(fat~ . ,meatspec , lambda = lambda_seq)

> select(ridgeregg)
modified HKB estimator is 2.363535e-08
modified L-W estimator is 0.907997
smallest value of GCV  at 3.25e-08

> matplot(ridgeregg$lambda, coef(ridgeregg) ,  xlab = expression(lambda) ,
ylab = expression(hat(beta)) ,col=1)
> library(Hmisc)
> minor.tick(nx=20 , ny = 10)
```

The R-code helps us to get the appropriate value for λ, which here is $3.25e-08$. A ridge trace is a plot between the values of β and the values of λ. A lambda where the trace seems to stabilize is used in ridge estimator. The ridge trace for this data is given in 3.2. It can be observed that the ridge trace seems to stabilize between $3e-08$ and $4e-08$

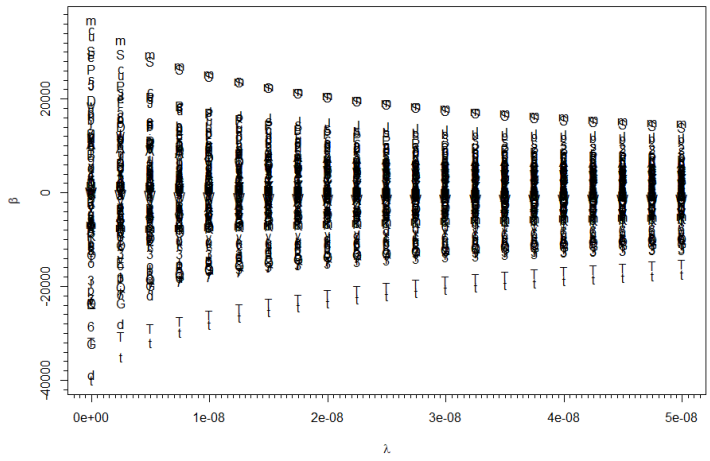


Figure 3.2: Ridge trace for “meatspec” dataset

Chapter 4

Density estimation and Smoothing

The sources used for this chapter are [15], [13], [12] , [5] , [6] , [11] and [3].

4.1 Density Estimation

The methods of estimating the distribution of the population from which the data is sampled is called **Density Estimation**. Here, we assume that the data comes from a continuous distribution. In a parametric setting density function relies on its parameters. Hence to estimate the density, it is enough to estimate the parameters involved in the density function. But these estimates of density will be valid only when data actually follows the assumed distribution.

It is very likely that someone who wants to know the distribution, does not know ahead of time that the density function belongs to a certain class of parametric distribution functions. In such cases, we need non-parametric estimation

4.1.1 Estimating CDF

Assume that X follows a continuous distribution f with CDF F, then for $h > 0$

$$P\left(x - \frac{h}{2} < X < x + \frac{h}{2}\right) = \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y)\mathrm{d}y \tag{4.1}$$

Now if f is smooth and h is small, then :

$$\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y)\mathrm{d}y \approx h \cdot f(x) \Rightarrow \hat{f}(x) = \frac{F(x+\frac{h}{2}) - F(x-\frac{h}{2})}{h}$$

Hence we have now moved to the problem of estimating CDF of an unknown distribution. The following are some of the ways of estimating cumulative density of an unknown continuous distribution.

Empirical CDF

If $X_1 \cdots X_n$ are IID random variables with distribution function F, then the empirical CDF is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{4.2}$$

Note that here we approximate the unknown distribution with a discrete distribution function. It is a step function estimate for F. From (4.1) and (4.2), we can show that

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} I(x - \frac{h}{2} \leq X_i \leq x + \frac{h}{2})}{nh} \tag{4.3}$$

Some properties of empirical density function (EDF) which can be shown easily are:

1. EDF $\hat{F}_n$ is an unbiased estimate of F
2. $MSE = Var(\hat{F}_n) = \frac{1}{n}F(x)(1-F(x))$

   Proof.

   $$Var(\hat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n}\left[E(\{I(X_i \leq x)\}^2) - \{E(X_i \leq x)\}^2\right] \tag{4.4}$$

   where $E(X_i \leq x) = F(x)$ and $E(\{I(X_i \leq x)\}^2) = F(x)$ and hence this proves the result. □

3. $\hat{F_n(x)} \longrightarrow F(x)$ in probability.

   Proof. This result can be proved use the chebyshev's inequality : $P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$ □

   In fact, it can be shown that $\hat{F_n(x)} \longrightarrow F(x)$ almost surely and uniformly.

Drawbacks: There are two main drawbacks for this estimate which have to taken seriously. They are:

1. A continuous distribution is being defined by a discrete non-continuous function.
2. If our data set is small, then we cannot expect a good estimate. In order to get a good estimate in this case, we need a large number of data points.

The following R-code shows how to plot an ecdf in R.

```
> data <- rnorm(50 , 0 , 1) ## store 50 normal random variables in
 variable "data"
> F50 <- ecdf(data) #f stores the emperical distribution function
> plot.ecdf(F50)
> data1 <- rnorm(10)
> F10 <- ecdf(data1)
> plot.ecdf(F10)
```

4.1.2 Histogram and Centred histogram

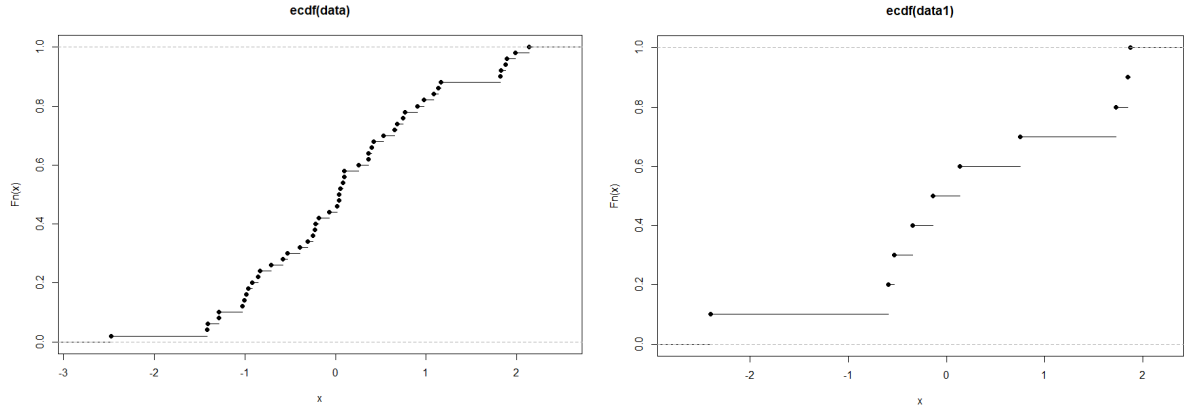
Histogram is a well known and popular density estimate for a continuous distribution. Formally, it can be defined in the following mathematical manner :

Let X be the random sample of size "n" from a continuous population then the histogram density estimate is given by

$$\hat{h}_n(x) = \frac{\text{no. of } X_i \in B_j}{nh} \text{ when } x \in B_j \tag{4.5}$$

where B_j are different partitions of $[0,1]$. Without loss of generality, we can assume that each X_i is in $[0,1]$, then

$$B_1 = \left[0, \frac{1}{M}\right); B_2 = \left[\frac{1}{M}, \frac{2}{M}\right) \cdots B_M = \left[\frac{M-1}{M}, 1\right) \text{ where } h = \frac{1}{M} \text{ is called bin width .}$$



(a) ECDF for 50 points drawn from normal distribution(b) ECDF for 10 points drawn from normal distribution

Figure 4.1: The distribution is approximated better with greater number of points.

It can be shown that the expectation of this estimate is:

$$E(\hat{h}_n(x))\frac{M}{n}\sum_{i=1}^{n} P(X_i \in B_j) = MP(X_i \in B_j) = M\left[F(\frac{j}{M}) - F(\frac{j-1}{M})\right] \tag{4.6}$$

Using the fact that $\frac{1}{M} = \frac{j}{M} - \frac{j-1}{M}$ and then applying mean value theorem to F, we get that there exists x^* such that

$$E(\hat{h}_n(x)) = \frac{\left[F(\frac{j}{M}) - F(\frac{j-1}{M})\right]}{\frac{j}{M} - \frac{j-1}{M}} = h(x^*);\ \text{for } x^* \in B_j \tag{4.7}$$

Applying mean value theorem to h, we get the existance of x^{**} such that

$$\frac{h(x^*) - h(x)}{x^* - x} = h'(x^{**}) \tag{4.8}$$

From (4.7) and (4.8), we can show that

$$Bais = E(\hat{h}_n(x) - h(x)) \leq \frac{\mid h'(x^{**}) \mid}{M} \tag{4.9}$$

$$Var(\hat{h}_n(x)) = MSE \leq \frac{(h'(x^{**}))^2}{M^2} + \frac{Mh(x^*)}{n} + \frac{(h(x^*))^2}{n} \tag{4.10}$$

Observation: Using similar calculations as shown above, we can show that From the inequalities in (4.9) and (4.10), we observe that as M increases (number of bins increases), the chance of over-estimating or underestimating the data reduces. Hence increasing the bin width increases the bias but reduces the variability. This is referred to as **over-smoothing**. Conversely if the bin width reduces, our bias reduces but variability increases. This is referred to as **under-smoothing**.

Drawbacks: The drawbacks of using histogram as density estimate are:

1. The estimate of a continuous distribution is not continuous
2. The density estimator depends on width and end points of certain fixed intervals which are chosen beforehand.

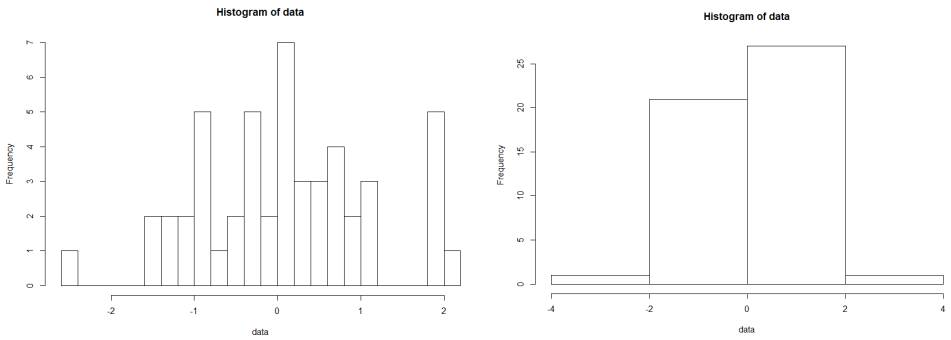
3. Increasing the number of bins might reduce the bias but it will also increase the chances of getting regions of 0 probability in the histogram.

The estimate's dependency on the bin width and its end points can be removed by defining a **centred histogram**. In a normal histogram, all B_j are fixed and pre-determined. But in the case of centred histogram, the band width is kept fixed at h but the intervals are not. The pdf estimate at x in this case is

$$\hat{h}_c(x) = \frac{\text{no. of } X_i \in \left(x - \frac{h}{2}, x + \frac{h}{2}\right)}{nh}$$

The R-codes below show how to plot a histogram and a centred histogram in R.

```
hist(data, breaks = 20) ## to get histogram with 20 bins, high variance,
0 prob region
hist(data, breaks = 3) #oversmoothed
```

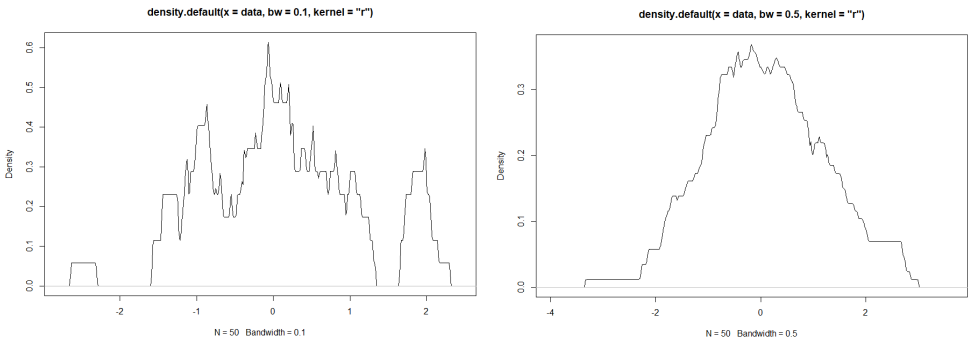


(a) High variance due to many bins (b) oversmoothed estimate

Figure 4.2: The distribution is not approximated well with too many or too few points.

Now, using the R-code below, we plot centred histograms and observe that increase in bin-width increases the smoothness of the estimate.

```
x <- density(data, kernel = "r", bw = 0.1)
z <- density(data, kernel = "r", bw = 0.5)
plot(x)
plot(z)
```



(a) bw = 0.1 (high variance) (b) bw = 0.5 (over-smoothed)

Figure 4.3: Plotting a centered histograms with bw = 0.1 and 0.5.

4.1.3 Kernel Density Estimates

A function K is called a kernel function if

- $K(x) \geq 0$
- $K(-x) = K(x)$
- $\int_{-\infty}^{\infty} K(x)\mathrm{d}x$

Example 4.1. Examples of some of the commonly used kernels are :

1. $K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ is called the Gaussian kernel
2. $K(x) = \frac{3}{4} max\{1 - x^2, 0\}$
3. $K(x) = \begin{cases} 1 & \text{if} -0.5 \leq x \leq 0.5 \\ 0 & \text{otherwise} \end{cases}$ is called the box-kernel

△

Disadvantages of histogram have provided the motivation for defining the **kernel density estimates**(KDE). Not only does kernel density estimate solve the problem of dependency on choice of bins but also solves the continuity problem i.e. KDE of a continuous function is continuous provided we choose a smooth kernel.

Remark 4.1. The order in which the estimates have developed is as given below:
Histogram $\longrightarrow$ Centred Histogram $\longrightarrow$ Kernel Density Estimate
Moving from histogram to KDE, we solved the problem of dependencies on bin and then we generalized the centred histogram to obtain KDE which solved the continuity problem as well.

Given $X_1, \cdots X_n$ and a kernel K, the KDE at point x is given by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right) \text{ where } h = \text{ bandwidth} \tag{4.11}$$

Theorem 4.1. *The definitions histogram, centred histogram and box kernel are all equivalent.*

Proof. Simple manipulation of the definitions can prove the above statement. □

Remark 4.2. From 4.1 above, we can view KDE as a generalization of a centred histogram. Changing the kernels give different estimates for the function.

Properties of KDE

Let $X_1 \cdots X_n$ be IID sample from an unknown population following a density function "$p(x)$". Initially, let us just consider a single point x_0. We analyse the quality of the estimate :

$$\hat{p}_n(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right)$$

Bais of the KDE

Using the basic definition of expectation (using integral) of function of random variables, we can show that

$$E(\hat{p}_n(x_0) - p(x_0)) = \left\{ \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x - x_0}{h}\right) p(x) \mathrm{d}x \right\} - p(x_0) \tag{4.12}$$

Applying the change of variable $y = \frac{x-x_0}{h}$ in the (4.12) and using the following Taylor expansion for p will help us simplify (4.12).

$$p(x_0 + hy) = p(x_0) - (hy)p'(x_0) + \frac{1}{2}(hy)^2 p''(x_0) + \mathcal{O}(h^2) \tag{4.13}$$

We use kernel properties, change of variable and (4.13) to simplify (4.12) to the folloing equation:

$$\text{Bais } = \frac{h^2}{2} p''(x_0) \int_{-\infty}^{\infty} K(y) y^2 \mathrm{d}y + \mathcal{O}(h^2) = \frac{1}{2} h^2 p''(x_0) \mu_k + \mathcal{O}(h^2) \tag{4.14}$$

μ_k in above line is a $\int_{-\infty}^{\infty} y^2 K(y) \mathrm{d}y$ From (4.14), we can say that if the bias is large then $p''(x_0)$ value is large $\implies$ there is more rate of change of slope $\implies$ the function p(x) will be more curved at x_0 (like peaks). Hence KDE tries to smoothen the peaks of a distribution function.

Variance of KDE

$$Var(\hat{p}_n(x_0)) = \frac{1}{n^2 h^2} \sum_{i=1}^{n} Var\left[K\left(\frac{X - X_i}{h}\right) \right] \leq \frac{1}{nh^2} \int_{-\infty}^{\infty} K\left(\frac{X - X_i}{h}\right) p(x) \mathrm{d}x \tag{4.15}$$

Again using the change of variable $y = \frac{x-x_0}{h}$ and applying (4.13) as before to (4.15) along with some kernel properties, we get can simplify (4.15) as follows:

$$Var(\hat{p}_n(x_0)) \leq \frac{1}{nh} p(x_0) \sigma_k^2 + \mathcal{O}(\frac{1}{nh}) \text{ where } \sigma_k^2 = \int_{-\infty}^{\infty} K^2(y) \mathrm{d}y \tag{4.16}$$

What we can tell from the inequality in (4.16) is that at a given point x_0 where density value $p(x_0)$ is large, the variance is also large.

$$\text{Now, } MSE = \left(\frac{h^2}{2} p''(x_0) \mu_k \right)^2 + \frac{1}{nh} p(x_0) \sigma_k^2 + \mathcal{O}(h^4) + \mathcal{O}(\frac{1}{nh}) \tag{4.17}$$

Then the term on the right handside in (4.17) is called **Asymptotic Mean Square Error(AMSE)**. Minimizing AMSE withrespect to the bandwidth h give us the following optimal value of $h_{opt}(x_0)$ for a given x_0.

$$h_{opt}(x_0) = \left[\frac{p(x_0 \sigma_k^2)}{n |p''(x_0)|^2 \mu_k^2} \right]^{\frac{1}{5}} \tag{4.18}$$

Remark 4.3. But the major problem is that the above optimal h is just a theoretical minima. We cannot use this in real life because we do not know the distribution $p(x)$.

In all the above analysis, we considered only one point x_0, but in general we want to control the overall MSE of the entire function. For one point x_0, $MSE = E[(\hat{p}_n(x_0) - p(x_0))^2]$. So for all points x, we have :

$$MISE = E[\int_{-\infty}^{infty} (\hat{p}_n(x_0) - p(x_0))^2 \mathrm{d}x] \tag{4.19}$$

$MISE$ is called the mean integrated square error. Now some calculations will let us show that

$$MISE(\hat{p_n}) = \left\{ \frac{1}{4} h^4 \mu_k^2 \int_{-\infty}^{\infty} |p''(x)| \mathrm{d}x \right\} + \frac{\sigma_k^2}{nh} + \mathcal{O}(h^4) + \mathcal{O}(\frac{1}{nh}) \tag{4.20}$$

The dominating term of (4.20) is called as the **asymptomatic mean integrated square error**. Then the optimal smoothing bandwidth can be obtained by minimizing $AMISE$ w.r.t h is given by:

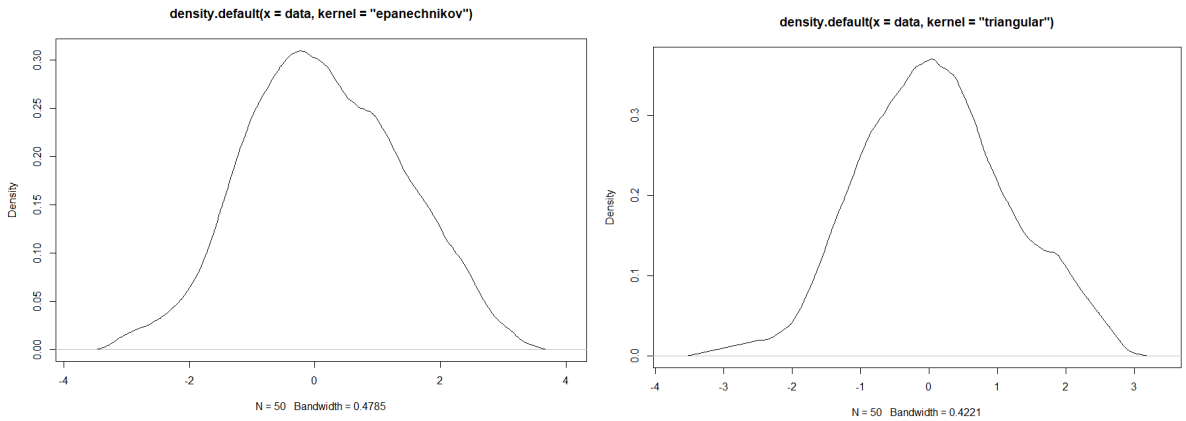
$$h_{opt} = \left[\frac{\sigma_k^2}{\mu_k^2 \left(\int_{-\infty}^{\infty} |p''(x)| \mathrm{d}x \right) n} \right]^{\frac{1}{5}} \tag{4.21}$$

Remark 4.4.

- Again (4.21) can't be used for practical purposes, because p is unknown.
- In-fact the problem about how to choose "h" is still unsolved and is known as **bandwidth selection problem**
- There are modifications to this procedure where we can use a variable band width. This has not been discussed here.

The following R-codes plot KDEs. We observe two things from these plots. First one is that choice of the kernel slightly effects the shape of the density estimate. This is shown in 4.4 Second one is that increasing the bandwidth increases the smoothness of the estimate.

```
plot(density(data, kernel = "gaussian"))
plot(density(data, kernel = "epanechnikov"))#smoother the kernel,
 smoother the estimate
plot(density(data, kernel = "triangular"))
plot(density(data, kernel = "gaussian", bw= 0.3 ))
plot(density(data, kernel = "gaussian", bw= 0.6 ))
```

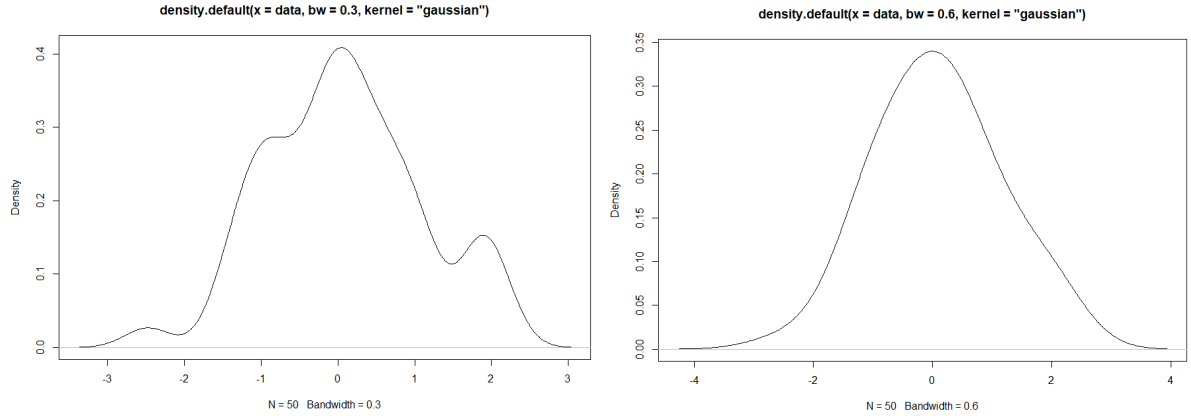


(a) KDE with kernel = epanechnikov kernel

(b) KDE with kernel = triangular kernel

Figure 4.4: Change in kernel does not effect the estimate so much

Now, we see how the KDE changes with change in bandwidth. It can be seen that increasing the bandwidth smoothens the data.



(a) Kernel = gaussian , bandwidth = 0.3 (b) Kernel = gaussian , bandwidth = 0.5

Figure 4.5: Increase in bandwidth smoothens the estimate

4.2 Smoothing

A smoothing algorithm is a summary of trend in Y as a function of $X_1 \cdots X_n$. Smoother takes the data and returns a function called **smooth**.In the bivariate case, a smoother is a procedure that is applied to the bivariate data $(x_1, y_1) \cdots (x_n, y_n)$ that produces a decomposition $y_i = s(x_i) + \epsilon_i$ where s is called the smooth function, also called as smooth.

In this chapter we discuss a couple of smoothing methods. But all smoothers considered will be linear. The following are the assumptions for this chapter:

1. There are n pairs of observations $(x_1, y_1) \cdots (x_n, y_n)$ and without loss of generality, assume that $x_1 \leq x_2 \leq \cdots \leq x_n$.
2. All observations are related through the expression $y_i = f(x_i) + \epsilon_i$ for $i = 1, 2, \cdots n$
3. ϵ_i's are IID froma continuous distribution centred at 0.

4.2.1 Local Averaging(Friedman)

This linear smoother is given by the below expression where x_j are such that $f(x_j) = y_j$ and x_j for $j = 1, 2 \cdots s$ are "s" points in the "neighbourhood" of x_i.

$$\hat{f}(x_i) = \frac{\sum_{i=1}^{s} y_j}{s}$$

Now, the neighbourhood of x_i is the smallest symmetric window about x_i containing s observations. The number of points in the window is called span. Note that the window size changes for for different values of x_i but *always* includes s data points. A larger span over-smooths the data and smaller spans provide an under-smoothed estimate. Friedman proposed using a **cross-validation** method to choose the span.

Cross-Validation

Let $\hat{g}_\lambda(x)$ be an estimate for $g(x)$. Then the **predictive squared error (pse)** of $\hat{g}_\lambda(x)$ is given by

$$pse(\lambda) = E[(y^* - \hat{g}_\lambda(x))^2]$$

where y^* is the **new** response value associated with the predictor x i.e. y^* and y_i are independent of each other for all i.

Cross-validation is an idea in regression where we estimate $pse(\lambda)$ by $CV(\lambda)$ where

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}_{\lambda,-i}(x_i))^2 \tag{4.22}$$

where $\hat{g}_{\lambda,-i}$ is the estimate for the g using the data points $x_1 \cdots x_{i-1}, x_{i+1}, \cdots x_n$. Eliminating the i^{th} point while estimating g makes sure that y_i is independent of all the other response values.

Choosing a span

Let $\hat{y}_{(i)}$ be the estimate of y_i determined by using all observed data points except (x_i, y_i). If s is the span, $e_i(s)$ is defined as $y_i - \hat{y}_{(i)}$, then s is chosen such that $\frac{\sum_{i=1}^{n} e_i(s)^2}{n}$ is minimized over a set S of possible span values. The span selected this way is called as the *global span* because this span is used for every point x_i. There are procedures for obtaining variable spans but that hasn't been discussed here.

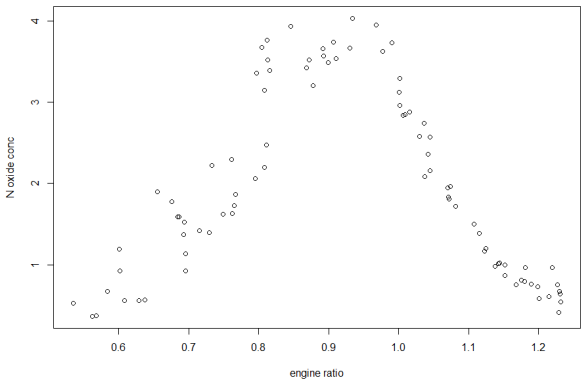
Now, we see an R-data example for the above theory. This data is about nitrogen oxide concentrations found in engine exhaust for ethanol engines. There are 88 pairs of data in this data set. We wish to smooth the data using "Friedman's local averaging".

```
> etoh <- lattice::ethanol  #ethanol data
> head(etoh)
    NOx  C     E
1 3.741 12 0.907
2 2.295 12 0.761
3 1.498 12 1.108
4 2.881 12 1.016
5 0.760 12 1.189
6 3.120  9 1.001

> plot(x = etoh$E , y = etoh$NOx , xlab = "engine ratio" ,
 ylab = "N oxide conc")
 ## data we want to smooth
#use "supsmu" to smooth the data using local averaging method
# use span = "cv" to use the cross validated variable span
> plot(supsmu(x = etoh$E, y = etoh$NOx , span = "cv"))
# cv span related smoothing might have a better balance of variance and bias
```

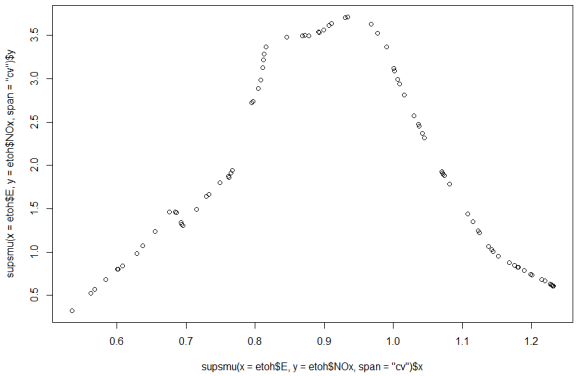
The data we wanted to smooth is given in Figure 4.6a while the Figure 4.6b has been smoothed using cross validated span.

```
#if span = p then it smooths the data using a constant span of size pn

> plot(supsmu(x = etoh$E, y = etoh$NOx , span = 0.05))
> plot(supsmu(x = etoh$E, y = etoh$NOx , span = 0.30))
# appears to be too smoothed . So biased
```

We observe how change in bandwidth changes how smoothed the data gets from 4.7. Clearly, increasing the span is increasing the smoothness and hence bias increases.

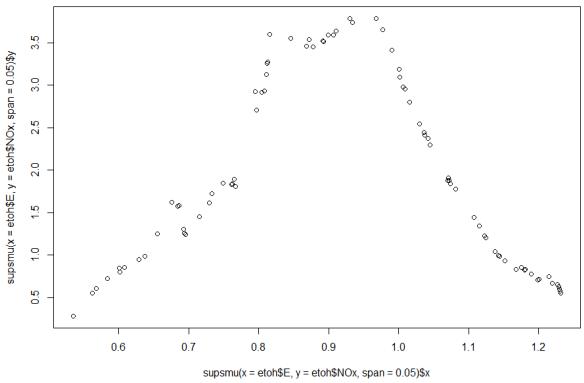


(a) The ethanol data we wish to smooth

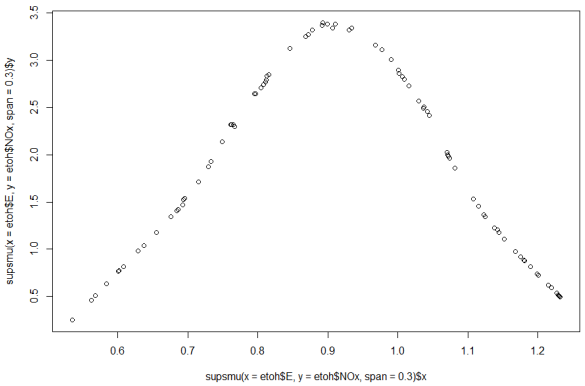


(b) Ethanol data smoothed using friedman's averaging and cv span

Figure 4.6



(a) Ethanol data smoothed using friedman's averaging and span =0.05



(b) Ethanol data smoothed using friedman's averaging and span =0.3

Figure 4.7

4.3 Kernel smoothing (Nadaraya and Watson)

Here we want to estimate a general function "f" rather than a density function. Nadaraya and Watson independently introduced the following kernel regression estimate:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} = \sum_{i=1}^{n} y_i w_i \text{ where } w_i = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} \tag{4.23}$$

Clearly, this estimate is linear in the observed data y_i and the weights w_i depend on kernel (K), bandwidth (h) and distance between x and x_i. Note that this is not a nearest neighbour method, unlike the previous two.

4.3.1 Deriving Nadaya-Watson estimator

Let (X_i, Y_i) be independent pairs of random variables such that we have $Y_i = m(X_i) + \epsilon_i$ given that $E(\epsilon_i | X_i = x) = 0$

$$\hat{m}(x) = E(Y|X = x) = \int y f_{Y|X}(y|x)\mathrm{d}y = \frac{\int y f_{X,Y}(x, y)\mathrm{d}y}{f_X(x)} \tag{4.24}$$

The marginal (f_X) and joint $(f_{X,Y})$ densities are estimated using the following kernel density estimates.

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \text{ and } \hat{f}_{X,Y}(x, y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) K\left(\frac{y-y_i}{h}\right) \tag{4.25}$$

Now use (4.25) in (4.24) and simplify in order to get the following

$$\hat{m}(x) = \frac{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \int_{\mathbb{R}} y K\left(\frac{y-y_i}{h}\right) \mathrm{d}y}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} \tag{4.26}$$

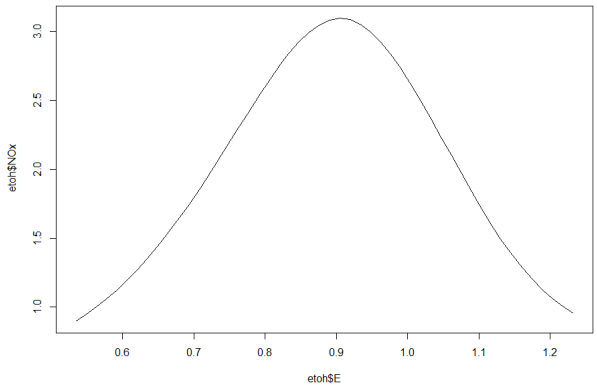
Now to simplify the integral, we use change of variable $\phi = \frac{y-Y_i}{h}$ to get

$$\int_{\mathbb{R}} y K\left(\frac{y-y_i}{h}\right) \mathrm{d}y = \int_{\mathbb{R}} h(h\phi + Y_i) K(\phi)\mathrm{d}\phi = \int_{\mathbb{R}} (h^2 + hY_i) K(\phi)\mathrm{d}\phi = 0 + hY_i \tag{4.27}$$

The step in (4.27) follows from kernel properties. Substituting (4.27) in (4.26) gives the desired result.

Now, we have R-codes using which we will see how to smooth the ethanol data using N-W estimator and a Gaussian kernel. The smoothed data is in 4.8a

```
# "npreg" command  implements the N-W kernel regression estimator
> etoh$NOx <- etoh$NOx[order(etoh$E)]
> etoh$E <- sort(etoh$E) # "npreg" requires x variable data to be sorted
> library(np)
> etoh.npreg <- npreg(bws = 0.09 , txdat = etoh$E , tydat = etoh$NOx)
> plot(etoh.npreg)
```



(a) Ethanol data smoothed using N-W estimator and a Gaussian kernel

Figure 4.8

Bibliography

[pen] Stat online. Lecture notes from different STAT courses. https://newonlinecourses.science.psu.edu/statprogram/. Last visited on 21 Nov 2019.

[2] (2013). Engineering statistics handbook. Online textbok. https://www.itl.nist.gov/div898/handbook/index.htm. Last visited on 21 Nov 2019.

[3] (http://www2.stat.duke.edu/ banks/218-lectures.dir/dmlect2.pdf). Smoothing. Lecture notes used at Duke University. http://faculty.washington.edu/yenchic/18W_425/Lec6_hist_KDE.pdf. Last visited on 21 Nov 2019.

[4] Buhlmann, P. and Machler, M. (Spring 2008). Computational statistics. Lecture notes from ETH Zurich. https://stat.ethz.ch/education/semesters/ss2012/CompStat/sk.pdf. Last visited on 21 Nov 2019.

[5] Castro, R. (2013). Introduction and the empirical cdf. Lecture notes used at Technische Universiteit Eindhoven , Netherlands. https://www.win.tue.nl/~rmcastro/AppStat2013/files/lecture1.pdf. Last visited on 21 Nov 2019.

[6] Chen, Y.-C. (2018). Density estimation: Histogram and kernel density estimator. Lecture notes for STAT 425 used at University of Washington. http://faculty.washington.edu/yenchic/18W_425/Lec6_hist_KDE.pdf. Last visited on 21 Nov 2019.

[Clarke] Clarke, B. R. *Linear Models*. Wiley series in probability and statistics. A Jhon Wiley and Sons publication, 2nd edition.

[8] Faraway, J. (2015). *Linear Models with R*. Texts in Statistical Science. CRC Press, Florida, 2 nd edition.

[Findsen and Troisi] Findsen, L. and Troisi, J. R tutorial for stat 350 lab 3. Lecture notes used at Purdue University . https://www.stat.purdue.edu/~lfindsen/stat350/Lab3R.pdf. Last visited on 21 Nov 2019.

[10] Friedman, Hastie, and Tibshiranir (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2 nd edition.

[11] Friedman, J. H. (1984). A variable span smoother. *Journal of American Statistical Association*, (1):1–16. https://www.slac.stanford.edu/pubs/slacpubs/3250/slac-pub-3477.pdf.

[12] Geyer, C. J. (2013). Smoothing. Lecture notes used at University of Minnesota. http://www.stat.umn.edu/geyer/5601/notes/smoo.pdf. Last visited on 21 Nov 2019.

[13] Hollander, M. (2014). *Non-parametric statistical methods*. Wiley series in probability and statistics. A Jhon Wiley and Sons publication, New Jersey, 3 rd edition.

[14] Li, P. (2005). Box-cox transformations : An overview. Lecture notes used at University of Connecticut . https://www.ime.usp.br/~abe/lista/pdfm9cJKUmFZp.pdf. Last visited on 21 Nov 2019.

[15] Montgomery, D. C. (2012). *Introduction to Linear Regression Analysis*. Wiley series in probability and statistics. A Jhon Wiley and Sons publication, New Jersey, 5 th edition.

[16] Ramachandran, K. and Tsokos, C. (2014). *Mathematical Statistics with Applications in R*. Academic Press, London, 2 nd edition.

[17] Seber, F. A. and Lee, A. J. (2003). *Linear Regression Analysis*. Wiley series in probability and statistics. A Jhon Wiley and Sons publication, New Jersey, 2 nd edition.

[Shalab] Shalab. Transformation and weighting to correct model inadequacies. Lecture notes for MTH 416 at IIT Kanpur . http://home.iitk.ac.in/~shalab/regression/Chapter5-Regression-TransformationAndWeightingToCorrectModelInadequacies.pdf. Last visited on 21 Nov 2019.

[19] Tanbakuchib, A. (2009). Assessing normality. Lecture notes for Pima Community College. http://www.u.arizona.edu/~kuchi/Courses/MAT167/Files/LH_LEC.0450.RandVars.AssesNorm.pdf. Last visited on 21 Nov 2019.

[Winner] Winner, L. Variance stabilizing transformations. Lecture notes from university of Florida. http://users.stat.ufl.edu/~winner/sta6207/transform.pdf. Last visited on 21 Nov 2019.