Contents

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Chapter 1

Simple Linear Regression

1.1 Introduction

One of the most fundamental aspects of statistical practice is to *analyze* and *interpret* the relationship between different variables in the population. What makes this interesting is that relationships or *association patterns* between variables are often as diverse as the variables themselves. Some variables may have a pretty simple relationship while others may have a more complicated pattern. Moreover, once the relationship between two variables has been determined, it is often of interest to "predict" the *unknown* value of one of those variables using the *known* value of the other. The branch of Statistics that deals with this problem is known as **Regression Analysis**.

Eg 1. Medical practitioners have long hypothesized that a person's muscle mass decreases with age. To explore this relationship in women, a nutritionist randomly selects 60 women between age groups 40 and 79 and calculates their muscle mass. Using the tools of regression analysis, you can help her figure out the "true" underlying relationship between age and muscle mass in the population of ALL women in that age range using the information from the above 60 women.

Eg 2. The crime rate of a region may depend on various factors like education (e), urbanization (u) and income (i) levels of that region. Using regression analysis, we can predict the crime rate of a particular region for given values of the above variables.

Eg 3. The price of houses/apartments in a particular city may depend on different factors like location (l), number of bedrooms (b), proximity to different attractions (metros, shopping malls) (a), etc. Using regression analysis, we can predict the price of a apartment *yet to be built* based on given values of the above factors.

The first step in any regression analysis exercise is to identify the concerned variables as *response* and *covariates* (or *explanatory variables*).

- **Response variable** (or Y) : This is the outcome or the dependent variable. In example 1, 2 and 3 above, , and are the response variables.
- **Explanatory variable** (or X) : This is the independent variable or the variable which *explains*

or *is related to* the outcome. In the above examples, , and are respectively the explanatory variables.

Note : *Whether a variable would be deemed response or explanatory often depends on the type of study. For example, if we want to know the relationship between muscle mass of a woman and the chance of her having osteoarthritis, muscle mass becomes the variable while having osteoarthritis (or not) is the .*

1.2 Plotting the data : Scatterplots

Generally the sample data will come as pairs i.e (X_1, Y_1), (X_2, Y_2),...,(X_n, Y_n) where each pair correspond to a unit and n is the sample size. Once the data is collected, it is advisable to plot those using a **scatterplot** to get a first hand visual impression of the association pattern between them. Every sample unit is represented by a point in the scatterplot. A scatterplot tells us

- How (and whether) the response and predictor variables are related to each other.
- If related, whether the relationship pattern can be reasonably approximated by a straight line.
- Whether there are any unusual points which falls well apart from the general trend of the points (outliers or influential points).

Figure 1.1. shows the scatterplot of muscle mass against age of the 60 women.

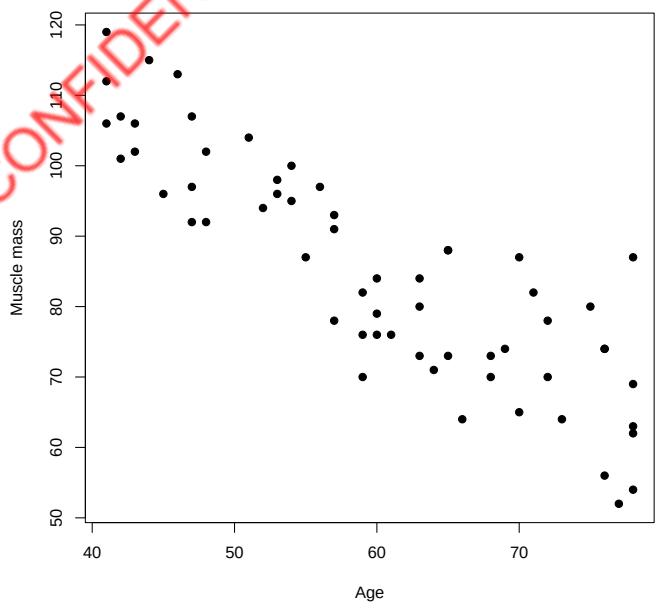


FIGURE 1.1: Plot of muscle mass against age of 60 women

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Observations :

1. The points show a strong trend i.e muscle mass and age seems to have a association. So, older women tend to have lower muscle mass on an average.
2. Age and muscle mass seems to have a linear relationship i.e the above trend can be approximated by a straight line reasonably well.
3. We do not see any point which falls well apart from the general trend of the points. i.e there does not seem to be any outliers or influential observations.

1.3 Correlation

When X and Y have an approximately linear relationship, we can actually go ahead and measure the strength of that relationship with a quantity called the **correlation coefficient**.

1.3.1 Population and Sample Correlation

There are two different quantities that might be called the correlation (or correlation coefficient) *vis*

- The **population correlation**, $\boldsymbol{\rho}$, measures the strength of the association (between X and Y) in the population.
- If we have a sample from a population, the **sample correlation** $\boldsymbol{r}$, measures the strength of the association in that sample. The formula for the sample correlation is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_X}\right)\left(\frac{Y_i - \bar{Y}}{S_Y}\right),$$

. Here $\bar{X}(\bar{Y})$ are the sample means of $X(Y)$ values while $S_X(S_Y)$ are the corresponding sample standard deviations [1]. Having said that, we will never actually calculate it by hand (any statistical software will do it for us). Naturally, we seldom know the value of ρ (since we never really have the population data), so we typically estimate it with the value of r.

1.3.2 Properties of the Correlation Coefficient

- The correlation coefficient always takes values between and .
- If X and Y have a positive association (as one goes up, the other tends to go up), then their correlation is positive.
- If X and Y have a negative association (as one goes up, the other tends to go down), then their correlation is negative.

[1] $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

- If X and Y have no linear association (in the population), then their population correlation ρ is zero. However, due to random variation, their sample correlation r will almost never be exactly zero but will be close to zero.
- *Correlation coefficient does not depend on the units of the variables nor on their identities (i.e response or explanatory)* - this is a big advantage of correlation coefficients.

The closer the correlation is to or , the stronger the linear association is between the two variables. For sample data, this means that the closer r is to -1 or 1, the closer the points on a scatterplot adhere to a straight line pattern, as shown in Figure 1.2.

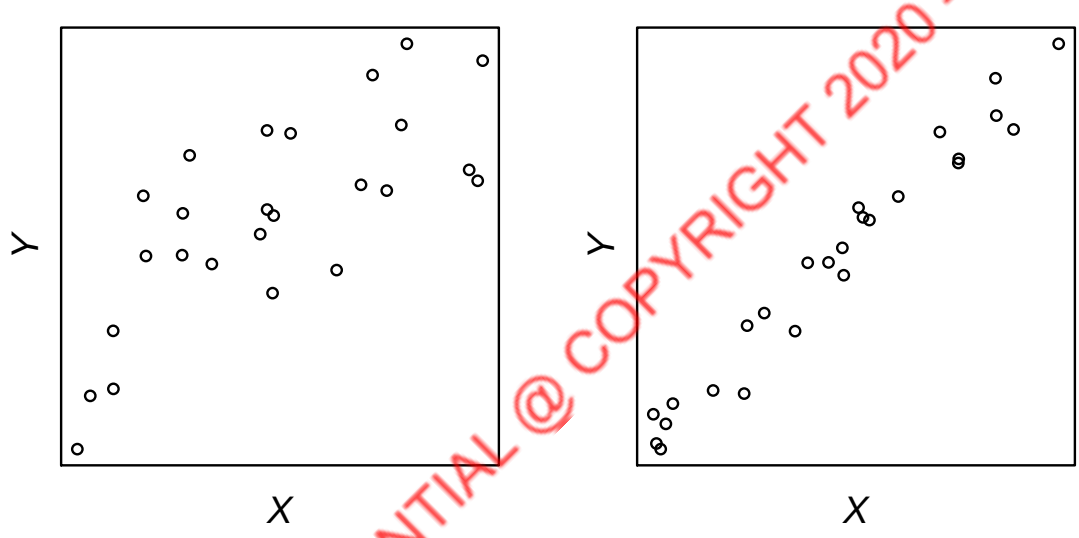


FIGURE 1.2: Visual interpretation of the correlation. The data sets on the left and right have $r = 0.75$ and $r = 0.98$, respectively.

For the age-muscle mass example, $r = -0.866$. So, we conclude that

- Since r is negative, age and muscle mass have a relationship i.e as age increases, muscle mass
- Since r is quite close to -1, we conclude that age and muscle mass have a *strong negative linear* relationship.

1.3.3 Caution : Linear Relationships only

The correlation is only useful for measuring relationships that are linear. Figure 1.3 shows two data sets both of which have r nearly zero. The scatterplot on the right clearly shows that X and Y have a non-linear (parabolic) relationship which cannot be quantified through the correlation coefficient. So, "$r = .06$" does not have a meaning in this context and can even be misleading. On the other

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

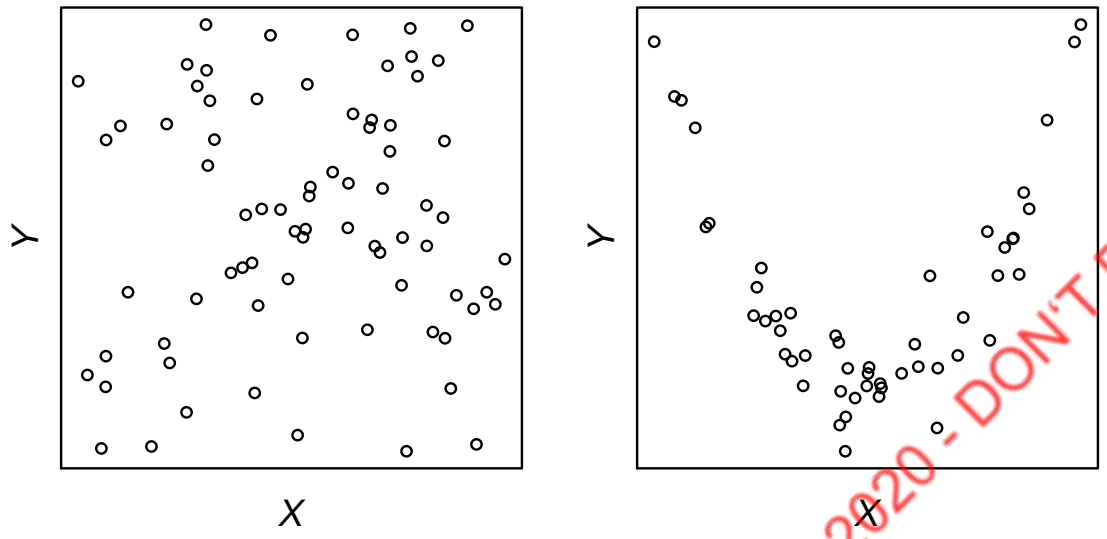


FIGURE 1.3: Two very different data sets with sample correlations near zero ($r = 0.07$ on the left, and $r = 0.06$ on the right).

hand, the fact that $r = .07$ for the first graph makes sense because it reflects the scatter of the points. Thus, in a nutshell, it is always a good idea to look at the scatterplot of the data first to see if the correlation is even a useful quantity to talk about.

1.3.4 Outliers

The presence of outliers can greatly influence the value of r. Figure 1.4 depicts it.

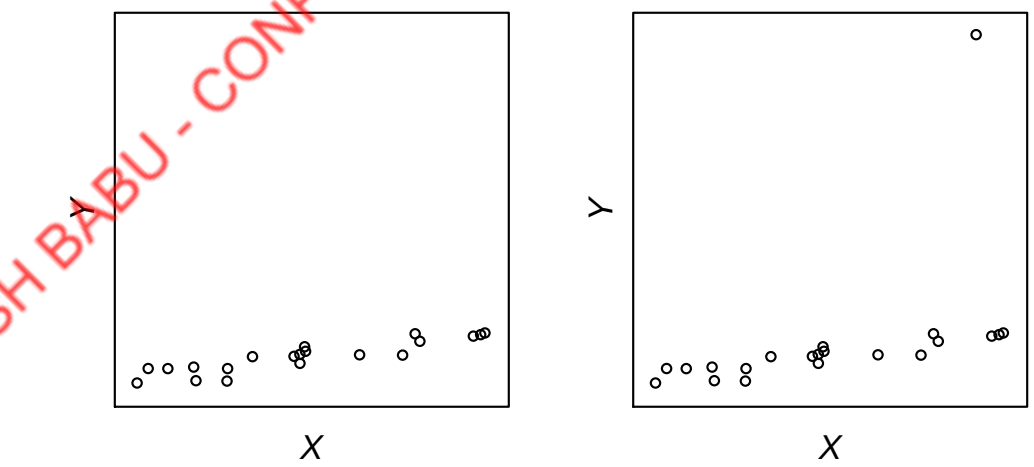


FIGURE 1.4: Effect of an outlier on the sample correlation. The data sets on the left and right have $r = 0.90$ and $r = 0.47$, respectively, despite differing by only a single observation.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Here the addition of one extra observation changes the sample correlation from $r = 0.90$ to $r = 0.47$. In fact, if the new point is a genuine outlier, the new pattern is not linear anymore. *Thus, correlation coefficient may not be meaningful when a dataset contains one or more outliers.*

1.4 Linear Regression Analysis

1.4.1 Concepts and Setup

When two variables, X and Y are linearly associated, we can go a step further by finding the equation of the straight line that best describes this pattern. Once this line is obtained, it can be used to predict an unknown value of the response variable (Y) from a known value of the explanatory variable (X). The procedure of doing this is known as **Linear Regression Analysis**.

To properly understand concepts about regression, we first need to understand how populations and samples relate to each other in the context of regression. Let us begin by considering a quantitative response variable Y and quantitative explanatory variable X. Each individual in the population has a value of X and a value of Y. It is easiest to think about the relationship between X and Y with an example.

Suppose, in the population of Indian adults, X is height (in inches), and Y is weight (in kilos). Let us consider only those people who are 65 inches tall ($X = 65$). Obviously all of them will not weigh the same, but their weights will vary around some mean value, which we will denote as, say $\mu_Y(65)$. (*We use μ to indicate the mean, Y to indicate what it is the mean of and* 65 *to indicate that it only refers to individuals with $X = 65$*) i.e $\mu_Y(65)$ is the population mean weight of all Indian adults whose height is 65 inches. Similarly the weights of all Indians who are 70 inches tall (i.e $X = 70$) will vary around some mean, say $\mu_Y(70)$, which we expect to be *greater* than $\mu_Y(65)$ (since height and weight are assumed to have a positive association).

Based on the above discussions, we are going to assume the following about the population :

- X and $\mu_Y(X)$ are related by a *straight line*.
- For each $X = x$, Y has a distribution about $\mu_Y(x)$.
- Each of the above distributions have the same

The above assumptions form the basis of linear regression analysis and is pictorially depicted below

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

The above straight line is expressed as

and is known as the population regression line. Here α is known as the while β is the. Since α and β are population parameters, we seldom know their actual values and will estimate those from the sample. Clearly,

- If Y and X have an increasing pattern (i.e $\mu_Y(x') > \mu_Y(x)$ for $x' > x$), then β
- If Y and X have a decreasing pattern (i.e $\mu_Y(x') < \mu_Y(x)$ for $x' > x$), then β
- If Y does not have a linear dependence on X, $\beta =$ i.e $\mu_Y(x) =$ regardless of the value of x.

1.4.2 Is the Relationship Linear?

There are many real-world situations in which the relationship between X and $\mu_Y(X)$ will *not* be linear. For example, Figure 1.5. plots 106 measurements of strontium isotopes found in fossil shells against their age.

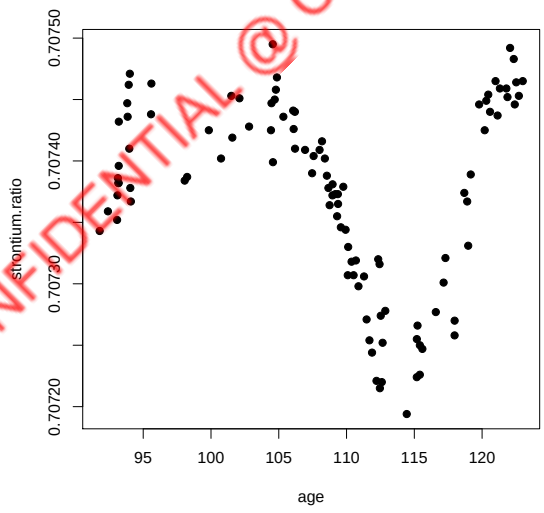


FIGURE 1.5: An example of non linear relationship

Clearly, it would be absurd to even try to fit a linear regression line to the above data. *So, it's important to think about whether the linearity assumption is at least somewhat sensible before deciding to conduct a study or analyze data using simple linear regression.* - this is where scatterplots come into play.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

1.4.3 Remember: Variability !!

Why do we need to write $\mu_Y(X)$ in (1) ? It's tempting to write something like

$$Y \overset{\text{NO!}}{=} \alpha + \beta X,$$

but this is unrealistic, because every individual with the same value of X *will not* have the same value of Y. (Does all PGPX students who study the same amount of time ends with the same CGPA/starting salary ? Does all companies having the same manpower generate the same revenue/year ?)

Instead, there will generally be some amount of variability in the values of Y for the same value of $X = x$. (Later on, we will make a further assumption that these Y values vary according to a certain distribution.)

1.5 Simple Linear Regression

This is the simplest (but also one of the most commonly used) form of regression analysis where our ultimate goal is to find the "best-fitting" straight line through a set of data points having a linear pattern.

1.5.1 Statistical Model

Let X and Y respectively be the explanatory and response variables. We have n pairs of data points vis (X_1, Y_1), (X_2, Y_2),...,(X_n, Y_n) where each pair correspond to a sample unit. X and Y are assumed to be related in the population as

$$\begin{aligned} Y &= \alpha + \beta X + \epsilon \\ &= \mu_Y(X) + \epsilon \end{aligned} \tag{1.1}$$

where ϵ are (unknown) error terms which are assumed to be independently and identically distributed (i.i.d) as $N(0, \sigma^2)$ (Normal with mean 0 and standard deviation σ^2) i.e for a particular value of $X = x$, Y is assumed to have a normal distribution with mean and variance . However, since we do not have population data, α and β are unknown and will estimated from the sample. This procedure, known as **Least Squares Estimation**, is done in such a way that the resulting straight line has the best possible fit to the given sample data.

1.5.2 Least Squares Estimation

In order to find the "best fitting" straight line through a given sample data say, (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n), we have to find the "best possible" estimates of α and β (say, $\hat{\alpha}$ and $\hat{\beta}$). We will use the sample data to get these estimates. The resulting line, also known as the *least squares regression line*, is the "best possible estimate" of the population regression line for the

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

given sample and is given by

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i, \qquad i = 1, 2, ..., n \tag{1.2}$$

where Y_i and $\hat{Y}_i$ are related by the equation

$$Y_i = \hat{Y}_i + e_i, \qquad i = 1, 2, ..., n \tag{1.3}$$

Here $(e_1, e_2, ..., e_n)$ are the *observed* errors and are known as the **residuals**. $\hat{\alpha}$ and $\hat{\beta}$ are obtained by minimizing the sum of squares of the above residuals i.e

$$\begin{aligned} S &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)^2 \end{aligned}$$

with respect to $\hat{\alpha}$ and $\hat{\beta}$. On doing so, we have

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{1.4}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \tag{1.5}$$

which are known as the **least squares estimates** of α and β.

Given $\hat{\alpha}$ and $\hat{\beta}$, the least squares regression line (or the estimated line) is given by

$$\hat{Y}(x) = \hat{\alpha} + \hat{\beta} x \tag{1.6}$$

where $\hat{Y}(x)$ is the *predicted value* of Y at $X = x$ while $\hat{\alpha}$ and $\hat{\beta}$ are the *sample* y - intercept and slope respectively.

The least squares regression equation 1.6 describes the $X - Y$ relationship in the sample, which will usually be close, but not exactly equal, to the relationship in the whole population, as shown in Figure 1.6. Thus, it makes sense to use the various parts of the regression equation, which we calculate based on the sample data, to estimate the corresponding parts of the equation that describes the population relationship, which we don't actually know.

1.5.3 Interpretation of Coefficients

$\hat{\alpha}$ and $\hat{\beta}$ are the sample Y-intercept and slope while α and β are their population analogues - these are called the regression coefficients. Since we will mainly deal with the estimated (or least squares) regression model, let us explore the meaning of $(\hat{\alpha}, \hat{\beta})$. The interpretation of (α, β) will be analogous in the population context.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

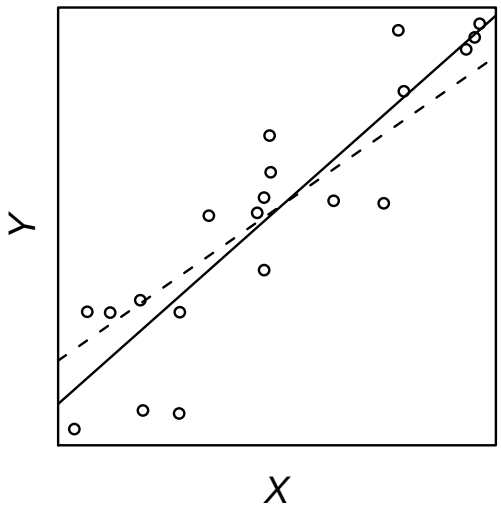


FIGURE 1.6: Regression equation (solid line) as an estimate of the unknown population relationship (dotted line).

Interpreting the slope

The slope tells us how a change in X affect a change in Y. Specifically, the sign of the slope indicates the direction of change while the value tells us the amount of change. Specifically,

- If $\hat{\beta}$ is positive, $\hat{Y}(X)$ with X which indicates that X and Y have a association.
- If $\hat{\beta}$ is negative, $\hat{Y}(X)$ with X which indicates that X and Y have a association.
- $\hat{\beta}$ represents the amount by which $\hat{Y}(X)$ changes when X is changed by

Note : *Unlike the correlation, the value of the slope will depend on the units in which X and Y are measured.*

Interpreting the y-intercept

The sample y-intercept, $\hat{\alpha}$ is the predicted value of Y when i.e $\hat{\alpha} =$

However, we have to be careful here. If $X = 0$ doesn't make sense, or if $X = 0$ is outside the range of our data (so that talking about what happens there would be extrapolation), then we *should not* interpret the y-intercept.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

1.5.4 Muscle Mass revisited

For the muscle mass example, the sample summary statistics are

$$\bar{X} = 59.98, \qquad S_X = 11.80, \qquad \bar{Y} = 84.97, \qquad S_Y = 16.21, \qquad r = -.866$$

Now, in order to calculate $\hat{\beta}$, we use the following alternative formulation

$$\hat{\beta} = r\left(\frac{S_Y}{S_X}\right)$$

Plugging in the values above, we have $\hat{\beta} =$ which in turn gives us $\hat{\alpha} =$

Thus, the least squares regression equation is given by

Figure 1.7. depicts the scatterplot of the age-muscle mass data with the above least squares regression line superimposed.

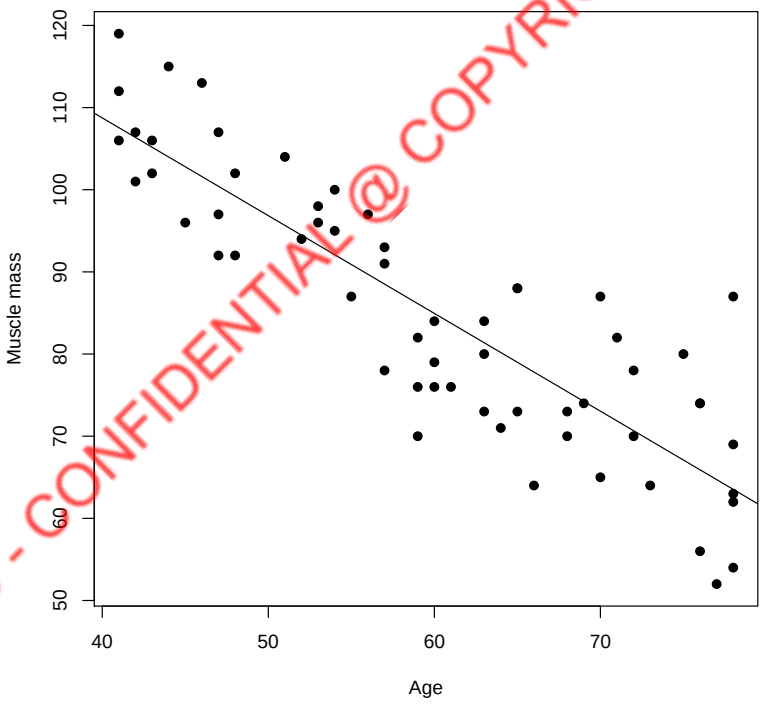


FIGURE 1.7: Least squares regression line fitted to the age-muscle mass data

Interpretation

- Since the slope is , muscle mass and age have a association. More specifically, for every 1 year increase in the age of a woman, her predicted muscle mass will decrease by .

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

- Since the y-intercept is , a woman with 0 yrs of age (i.e new born girl child) is predicted to have a muscle mass of . (Obviously this is absurd and a gross extrapolation - so, in this case, we *should not* interpret the intercept).

1.6 Making Predictions

https://www.overleaf.com/project One of the most important use of the least squares regression equation is to predict unknown values of the response variable (Y) from given or known values of the explanatory variables (X). We can predict the value of Y for any particular value of X by simply plugging that value of X into the regression equation and seeing what we get for $\hat{Y}$. However, we need to keep in mind that this X value should come from a subject who is similar to the subjects sampled in the original data set. Otherwise, we may run the risk of extrapolation.

If we are predicting for a in-sample subject, the prediction will not be exactly equal to the actual Y value of that subject. This is because of the variability that is inherent in our model (think of the ϵ_i's in the population regression model). All in all, the prediction is just our single-number best guess for a Y value at a particular X value. Figure 1.8. illustrates the idea of prediction.

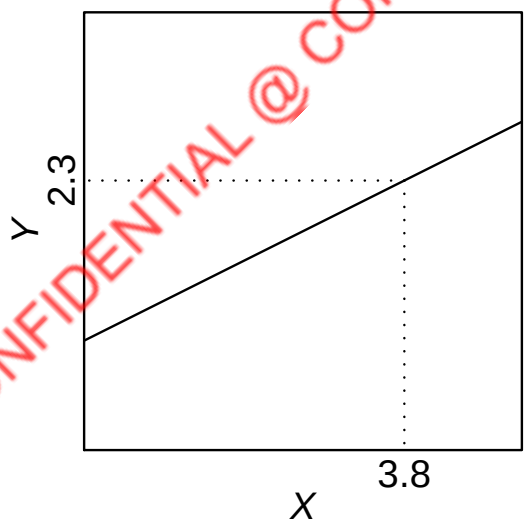


FIGURE 1.8: Visualization of a prediction using the regression equation (solid line). Here, the predicted value for $X = 3.8$ is $\hat{Y} = 2.3$.

One of the women sampled for the Muscle-mass data was Mrs. Tripathi who is 56 years old. So, her predicted muscle mass will be

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

1.6.1 Residuals

Let Y_i be the actual response value for an in-sample subject, say subject i while $\hat{Y}_i$ be the corresponding predicted value obtained by plugging in the corresponding X value in the least squares regression line. Then, the **residual** for subject i is given by

$$\text{Residual}_i = Y_i - \hat{Y}_i. \tag{1.7}$$

In this way, we can obtain the residuals of all the sampled subjects. Clearly, closer the residuals are to better will be the prediction. Suppose the actual muscle mass of Mrs. Tripathi is 97. Then her residual will be

1.6.2 Predictive Ability

As mentioned above, one of the most important use of the regression equation is to predict unknown values of the response variable (Y) from given or known values of the explanatory variables (X). Figure 1.9. illustrates regression equations those are good and bad at making predictions.

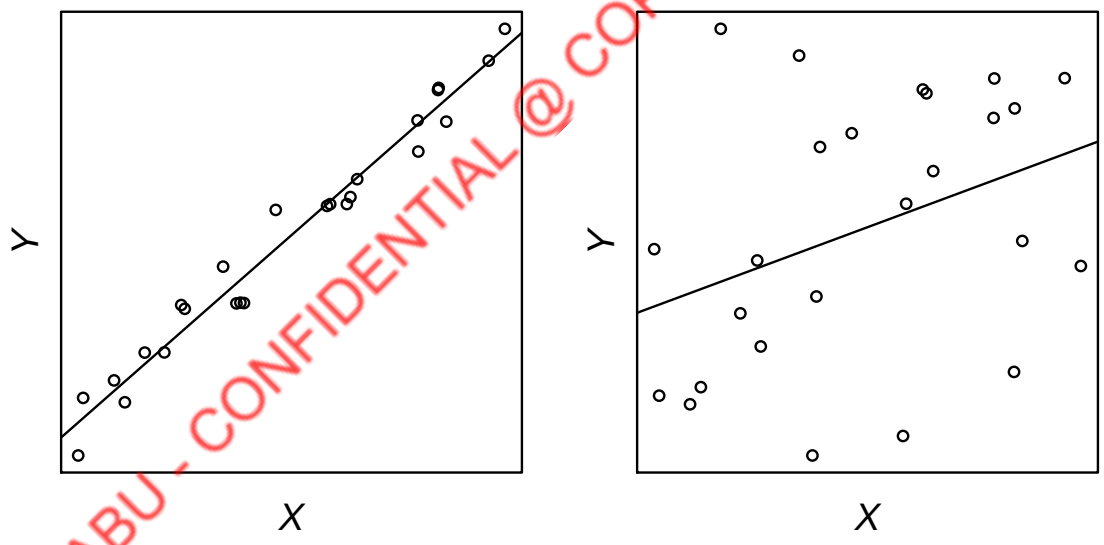


FIGURE 1.9: Visual interpretation of predictive ability. The regression on the left has better predictive ability than the regression on the right.

It is obvious that closer the observed values to the corresponding predicted values (on the line), better is the predictive ability of the least squares regression line.

1.6.3 Coefficient of Determination

We can quantify predictive ability using the coefficient of determination, R^2, which is just the square of the correlation coefficient r i.e $R^2 = r^2$. Clearly R^2 would range from

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Basically R^2 tells us how much better we are doing by regressing Y on X rather that just using to $\bar{Y}$ to predict Y.

- Better the predictive ability of our fitted regression model, closer R^2 is to
- Poorer the predictive ability of our fitted regression model, closer R^2 is to

We have already seen that the correlation coefficient of age and muscle mass is -0.866. Hence the coefficient of determination for the least squares regression will be

$$R^2 =$$

So, we conclude that the least squares regression line has a pretty good predictive ability since R^2 is quite high. Moreover, the above regression line results in 75% *less* error in predicting muscle mass compared to $\bar{Y}$.

R^2 can also be interpreted as the amount of variability in the response Y explained by the linear regression of Y on X. So for the above example, we can also say that of the variability in muscle mass can be explained by age. So, whichever way we see it, the least squares prediction equation does a pretty decent job in predicting muscle mass from age.

1.7 Outliers and their effect

Just as outliers can impact the correlation coefficient, they can greatly influence the regression line as well. Figure 1.10. illustrates the dramatic effect that a single outlier can have on the fitted regression line.

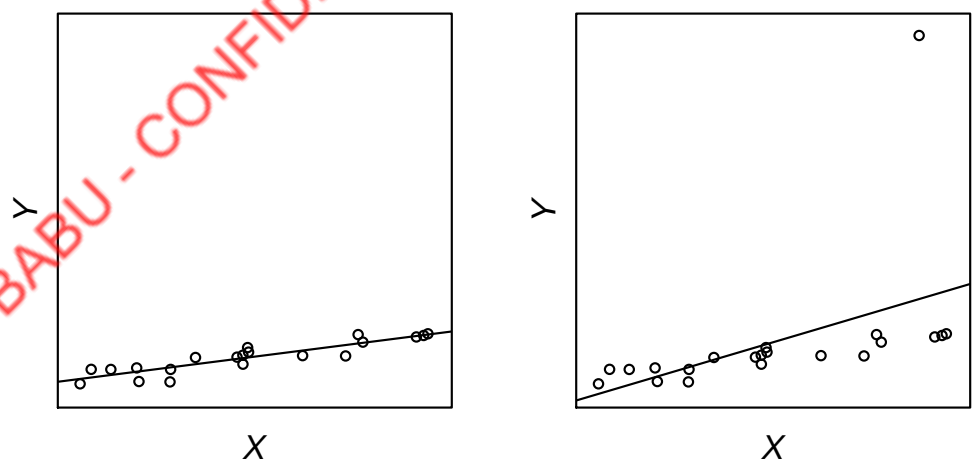


FIGURE 1.10: Effect of an outlier on the regression equation. The data sets yield regression equations that differ substantially despite differing by only a single observation.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

When the data contains one or more outliers, the regression equation can fit the data very poorly, and any results we obtain might be unreliable. So it's always a good idea to look at a scatterplot to make sure there are no outliers when doing simple linear regression.

1.8 Extrapolation

As mentioned in Sec 5.5, when should avoid making predictions for a new subject (i.e out-of-sample), whose X values are outside the range of the sample values i.e significantly smaller than the smallest X value or significantly larger than the largest X value in the sample. Making a prediction at an X value which is significantly outside the range of the X values in the data is called **Extrapolation**, and it leads to predictions that are unreliable or even ridiculous.

If we have daily temperature data of Ahmedabad from 1984-2015, making a prediction for 2016 would technically be extrapolation, but it might be okay for most practical situations. However, we probably would not want to use that data to make a prediction for 2300.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

1.9 R codes

Following R codes were used to perform the analysis in this chapter. Necessary explanations are provided alongside the codes. (R can be downloaded from `https://www.r-project.org/`)

Important : *Create a folder, say* `SDA` *in your desktop. Transfer all the datafiles therein. Open the folder, right-click on it and go to* `Properties` *and copy the* `Location` *indicator, say* `D:Desktop`. *Now you need to add the folder name to it and change the backslashes to forward slashes and use this path name in the codes below, as I have done. Remember, your pathname will be unique and may be different from* `D:/Desktop/SDA`.

1. Setting up the working directory :
```
setwd("D:/Desktop/SDA")
getwd()
```

2. Plotting the `age-muscle mass` data, Figure 10.1 :
```
musclemass<-read.table("D:/Desktop/SDA/musclemass.txt",header=T)
attach(musclemass)
plot(X,Y,xlab="Age",ylab="Muscle mass",pch=19)
```

3. Saving the above picture as a `.pdf` file in the working directory :
```
pdf("Musclemass.pdf")
plot(X,Y,xlab="Age",ylab="Muscle mass",pch=19)
dev.off().
```

4. Correlation coefficient of Age and Muscle-mass :
```
r(X,Y).
```

5. Plotting the `fossil` data, Figure 10.5 :
```
fossil<-read.table("D:/Desktop/SDA/fossil.txt",header=T)
attach(fossil)
plot(age,strontium.ratio,xlab="Age", ylab="Strontium.ratio",pch=19)
```

6. Fitting the least squares model to age-musclemass data (Sec 10.5.2) :
```
fit.mm<-lm(Y~X)
summary(fit.mm).
```

7. Figure 10.7 :
```
fit.mm<-lm(Y~X)
abline(fit.mm)
```

8. Obtaining fitted values and residuals for sampled subjects (Sec 10.6) :
```
fitted.mm<-fitted(fit.mm)
resid.mm<-residuals(fit.mm)
```


DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Chapter 2

Parameter Estimation

2.1 Motivation

One of the most basic questions we should address in any regression analysis problem is whether Y and X are linearly associated. Specifically in a linear regression problem, this question translates into checking whether or not . This is because, if $\beta = 0$, $\mu_Y(X) =$ which implies that Y and X have no linear association between them. *For instance, in the age-muscle mass example, $\beta = 0$ would imply that muscle mass and age are not linearly related in the population.*

There are two distinct ways in making inferences about parameters in general *viz* **Hypotheses tests** and **Confidence intervals**. These will be discussed next.

2.2 Hypotheses Tests for β

Here we will learn to perform a **Regression t test** for β. For that, we need to go through the following steps :

2.2.1 Hypotheses Formulation

For testing whether Y and X are linearly associated, we have the following two hypotheses with respect to β

$$\begin{aligned} \textit{Null hypotheses} &: H_0 : \beta = 0 \Rightarrow \text{Y and X have no linear association.} \\ \textit{Alternative hypotheses} &: H_a : \beta \neq 0 \Rightarrow \text{Y and X are linearly related.} \end{aligned}$$

The above alternative hypotheses is a two-sided one. Depending on situation, we can also use one sided alternative hypotheses $H_a : \beta > 0$ (i.e linear association) or $H_a : \beta < 0$ (i.e linear association). For instance, in the age-muscle mass example, we might want to test $H_a :$ since age and muscle mass generally have a negative association.

2.2.2 Test statistic

The test statistic for testing the above hypotheses is given by

$$t = \frac{\hat{\beta} - 0}{\hat{se}(\hat{\beta})} \sim t_{n-2} \quad \text{under} \quad H_0 \tag{2.1}$$

where $\hat{se}(\hat{\beta})$ is the estimated standard error of $\hat{\beta}$. Any statistical software will give us the value of the above test statistic.

2.2.3 P-values

Given the alternative hypotheses and the observed value of the test statistic above, we will be able to calculate the p-value as explained in Chapter 7. Still, a brief explanation is provided below for recapitulation.

Definition 1. *The p-value is the probability of getting a test statistic value at least as extreme as the one observed, assuming H_0 is true.*

For calculating the p-value, we need to keep the alternative hypotheses in mind. Suppose we observe a test statistic value of say, $t = 2.2$ for a t distribution with 30 degrees of freedom. If the alternative hypotheses is of a $>$ ($<$) type, then the p-value will be the area above (below) 2.2 under a t curve with 30 degrees of freedom. However, if the alternative is two-sided (i.e $\neq$), the p-value will be the combined area above 2.2 and below -2.2 for a t distribution with 30 degrees of freedom. This is represented graphically in Figure 2.1.

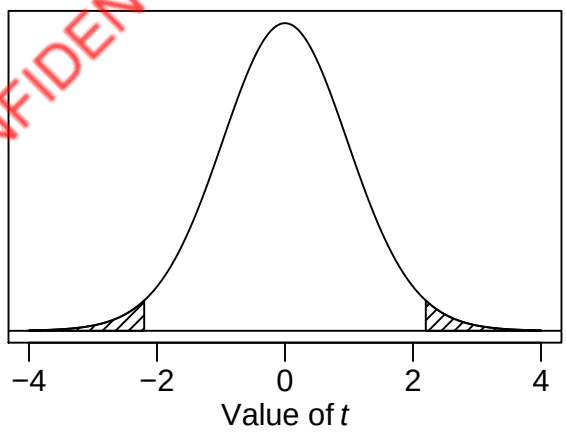


FIGURE 2.1: Two-tailed probability of a t distribution, $df = 30$.

Since the t distribution is symmetric, we typically find the probability for just one of these tails, usually the one on the right, and then double it to get the p-value. Here is a quick review of some of the properties of the t distribution:

- It is symmetric and centered at zero, with both a positive and a negative tail.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

- Its exact shape is determined by its degrees of freedom.
- Although it looks like a standard normal distribution, a t distribution has thicker tails than the normal. However, as the degrees of freedom gets larger, the t gets closer to a standard normal.

If we do not have access to statistical software to calculate p-values for us, we often have to use a t table like the one in the back of our textbook to try to figure out the p-value. A typical t table, like the one shown in Figure 2.2 below. has rows corresponding to different df values. Within the appropriate row, the table shows the test statistic values that correspond to certain right tail probabilities. We can use this information to figure out an approximate right tail probability for any test statistic value we want, and we then double this to obtain the p-value for a two-sided test. **Remember, software outputs generally provides the two-sided p-values.**

	Right-Tail Probability					
$df \downarrow$	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
⋮	⋮	⋮	⋮	⋮	⋮	⋮

FIGURE 2.2: First few rows of a t table.

Larger the p-value, is the evidence against H_0 i.e we are more likely to reject H_0 for p-values.

2.2.4 Decision

Finally, we need to make a decision about whether H_0 is reasonable, or whether we have enough evidence to reject H_0 and believe H_a instead. We do this by comparing the p-value to our chosen significance level α (often 0.05) and making a decision the same way we always do:

- If the p-value is less than or equal to α, we H_0 at α.
- If the p-value is greater than α, we H_0 at α and conclude that we do not have enough evidence to reject H_0.

In the context of our actual hypotheses (with a two-sided alternative), this means the following:

- Reject H_0 $\Rightarrow$ significant evidence (at α) to believe that Y is linearly associated with (or dependent on) X.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

- Fail to reject $H_0 \Rightarrow$ do not have significant evidence (at α) to believe that Y linearly depends on X i.e no point in trying to use X to predict Y.

Note : *If $H_a : \beta > 0$ and we fail to reject H_0, it may not imply that Y and X are independent. It may so happen that $\beta < 0$ i.e Y and X have a negative association. Hence, we can only conclude that, given the data, there is strong evidence that Y and X does not have a positive association. This is why, we should not say "we accept H_0".*

2.2.5 Age-muscle mass revisited

Let us go through all the steps of a two-sided regression t-test for the age-muscle mass example.

Assumptions

- 60 women were selected randomly satisfying the random sampling assumption. ✓
- One pair of observations for each women implying independence. ✓
- Mean muscle mass and age are linearly associated. (vide Fig 6.1). ✓
- Muscle mass have a normal distribution in the population (to be shown later). ✓
- Muscle mass have a constant spread for different age values (vide Fig 6.1). ✓

Hypotheses

H_0 : $\beta = 0$ *i.e* muscle mass and age have no linear association.

H_a : $\beta \neq 0$ *i.e* muscle mass and age are linearly related.

Test Statistic

The `R` output for the muscle mass data is given below

Predictor	Estimate	St. Error	t-value	$P(>\|t\|)$
Intercept	156.35	5.51	28.36	<2e-16 ***
Age	-1.19	0.0902		

TABLE 2.1: Parameter estimates for muscle mass data.

Thus, the test statistic would be

Since the sample size is 60, the degrees of freedom (df) of the above statistic will be

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

P-value

Since the alternative hypotheses is sided, the p-value will be the area under a t curve with df. The relevant t scores for this degrees of freedom are as follows :

	Right-Tail Probability					
$df \downarrow$	0.100	0.050	0.025	0.010	0.005	0.001
58	1.296	1.672	2.002	2.392	2.663	3.237

TABLE 2.2: Right tail probabilities for t_{58}

So, the p-value would be

Decision

Since our p value is extremely low (approximately 0), we H_0 and conclude that there is strong evidence of linear association between age and muscle mass. Figure 2.3 shows the t test in a nutshell.

t value "far" from 0	t value "near" 0
↓	↓
Small p-value	Large p-value
↓	↓
Evidence against H_0 (for H_a)	No evidence against H_0 (for H_a)
↓	↓
Reject H_0	Fail to reject H_0
↓	↓
Evidence that Y depends on X	No evidence that Y depends on X

FIGURE 2.3: Results and interpretations of a (two-sided) regression t test.

Note : *If our alternative was* $H_a : \beta < 0$ *instead, the one sided p-value would also had been* ≈ 0 *(since the t statistic is negative), we still would have rejected* H_0 *and conclude that there is strong evidence to believe that mileage and weight are negatively associated.*

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

2.3 Confidence Interval for β

Hypotheses tests simply tells us whether it is reasonable that $\beta \neq 0$. Instead, it might be interesting to figure out the set of all reasonable values of β. We achieve this by constructing a confidence interval of β.

The assumptions required to construct a confidence interval for β are exactly the same as those used for the regression t test which we have already discussed.

2.3.1 Formula

A $100(1-\alpha)\%$ confidence interval for β is given by

$$\left(\hat{\beta} - t_{\alpha/2,n-2}\hat{se}(\hat{\beta}), \hat{\beta} + \hat{se}(\hat{\beta})t_{\alpha/2,n-2}\right). \tag{2.2}$$

where the standard error of $\hat{\beta}$ is the same as the one used in the t statistic. The t-score is a number from the t table that depends on two things:

- It depends on the desired confidence level. Higher confidence levels require intervals, which mean larger t-scores. 95% is the most commonly chosen confidence level.
- It depends on the degrees of freedom. For simple linear regression, the degrees of freedom is $n-2$ since there are 2 parameters (α and β) to estimate. For a given confidence level, the t-score decreases as the degrees of freedom i.e the confidence interval gets narrower (hence, more precise) as the sample size

Some t tables, including the one we will use, provide a second set of column headings called "Confidence Level," as shown in Figure 2.4. Simply find the row for the appropriate *df* and the column for the appropriate confidence level, and the number in the body of the table is the t-score that should be used in constructing the confidence interval.

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df ↓	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
⋮	⋮	⋮	⋮	⋮	⋮	⋮

FIGURE 2.4: First few rows of a t table, with headings for confidence level.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

2.3.2 Interpretation

The standard interpretation of a confidence interval (let's say 95%) for β is that we are 95% confident that the true value of β is between (lower number) and (higher number). Loosely speaking, it sometimes helps to just think of a confidence interval for β as the set of possible values of β that are reasonable based on the data. We can fine-tune what we mean by "reasonable" by adjusting the confidence level.

2.3.3 Age-muscle mass revisited

From the hypotheses test we concluded that there is strong evidence of a linear association between age and muscle mass i.e $\beta \neq 0$. Let us now figure out the reasonable values of β by calculating a 95% confidence interval of the same.

From Table 2. we have $t_{0.025,58} \approx 2.00$ and $se(\hat{\beta}) = .0902$. So, a 95% confidence interval of β would be

i.e as age increases by 1 year, the average muscle mass will by at most and at least . Thus, both the two sided significance test and the confidence interval gives us the same conclusion regarding the slope.

Note : *We rarely perform inferential procedures for α because often its interpretation is not realistic. However, hypotheses tests and confidence interval procedures for α work exactly the same way as those for β.*

2.4 R codes

Following R codes were used to perform the analysis in this chapter. Necessary explanations are provided alongside the codes.

1. Setting up the working directory :

```
setwd("D:/Desktop/SDA")
getwd()
```

2. Table 11.1 (Sec 11.2.5) :

```
fit.mm<-lm(Y~X)
summary(fit.mm)
```


DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Chapter 3

Regression Diagnostics

3.1 Introduction

In fitting a linear regression model to a given data set, we have to make the following assumptions.

1. The sample has been selected through simple random sampling.
2. Observations corresponding to the sample units are independent of each other.
3. Y (or $\mu_Y(X)$) has a linear association with X in the population.
4. Y values corresponding to any particular X value has a normal distribution in the population.
5. Y values corresponding to any particular X value has the same spread (or standard deviation).

Now, some (or all) of these assumptions may not hold for a particular data set. In that case, it will be fallacious to use a linear regression model to draw inferences about that data. **Regression Diagnostics** refers to the procedure of checking whether a linear regression model is appropriate for a particular dataset (in the sense that the model satisfies the assumptions on which it is based). This is achieved through a procedure called **Residual Analysis**.

3.2 Residual Analysis

It turns out that if a regression model fails to satisfy some assumptions for a given data set, it gets reflected very clearly in the residuals of the fitted model. So, an examination of the residuals is a very effective way of checking the appropriateness of a regression model for a particular data set. This is implemented by plotting the residuals or standardized residuals (an improved version of the residuals) against the covariate/s (X) or the fitted/predicted values ($\hat{Y}$).

For example, one of our assumptions for the population regression model is that $\epsilon_i \sim N(0, \sigma^2)$ (where ϵ_i's are the errors). If a regression equation fits the data well, the residuals e_i's should also tend to be independently distributed about 0 with constant variance σ^2. Thus, an examination of

the residuals should give us a pretty good idea whether the above assumption has been satisfied by the regression model. This is the basis for residual analysis.

3.3 Type of Departures

The first two assumptions viz. randomization and independence of the observations can be hard to check once the data has been collected. So, it is important to design studies/surveys carefully to ensure that these two assumptions are valid. However, residual analysis can be used to check the last three assumptions namely linearity, normality and homoscedasticity as detailed below :

3.3.1 Non-linearity

When Y and X have a linear pattern *vis-a-vis* a linear regression model is appropriate for the data, the residuals, *when plotted against* X, tend to be randomly scattered above and below 0 having no particular pattern. Figure 3.1. shows the residual plot for the age-muscle mass data.

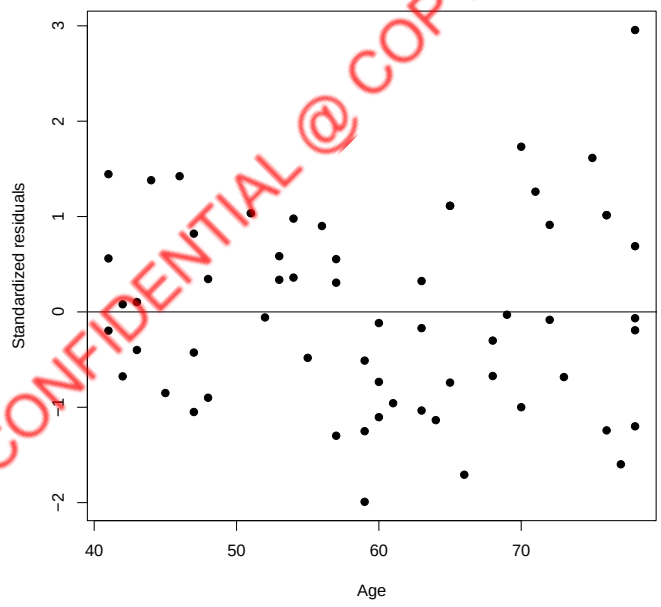
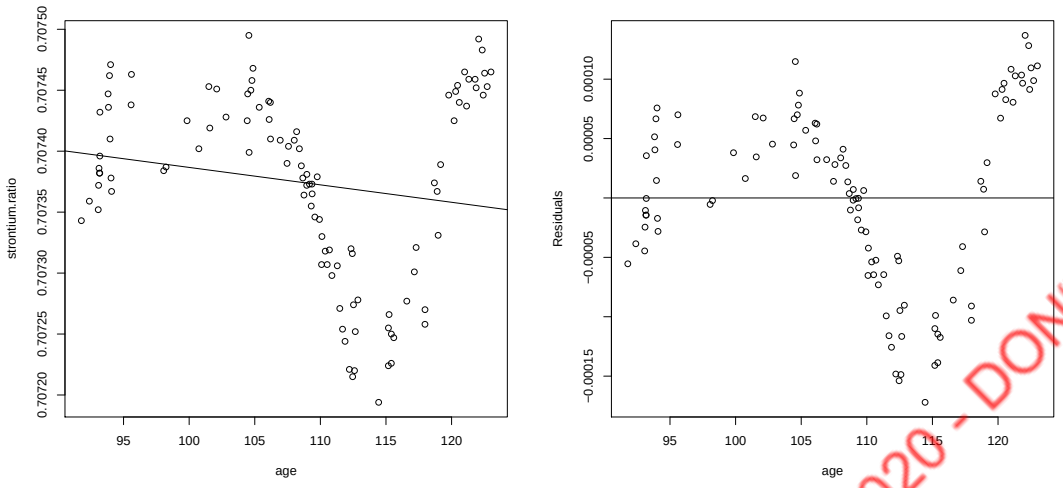


FIGURE 3.1: Residual plot for the age-muscle mass data.

Clearly, the residuals roughly follow a random pattern above and below the 0-line. Thus the linear regression model seems to be appropriate for this data set.

Figure 3.2. shows the non-linear fossil data (with the fitted least squares regression line) and the corresponding residual plot against X (age).

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE



(a) Fossil data (b) Residual plot

FIGURE 3.2: Nonlinear data and accompanying residual plot.

It is clear that the residuals follow a systematic (non-random) pattern about 0 indicating that a straight line fit is not at all appropriate for this data.

Thus, appropriateness of a linear regression model (for a dataset) will be indicated by a random scattering of residuals about 0 while inappropriateness of a linear fit will be indicated by a systematic (non-random) pattern of residuals about 0.

3.3.2 Non-constant error variance

A residual plot also indicates whether the assumption of constant error variance ($V(\epsilon_i) = \sigma^2$) has been satisfied. We have the following rule of thumb :

- If the error variance is constant, the residuals will be randomly scattered about 0.

- If the error variance increases with X, the residuals will have spread as X increases.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

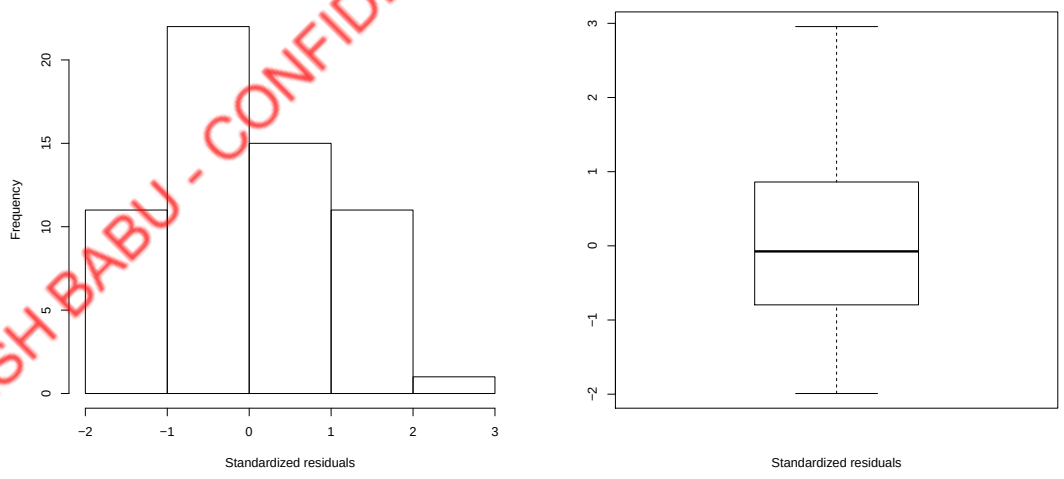
- If the error variance decreases with X, the residuals will have spread as X increases.

In Figure 3.1. we do not see any particular increasing or decreasing pattern of the residuals with age (X). This implies that the error variance may be independent of X and hence constant hence satisfying the *homoscedasticity* assumption.

3.3.3 Non-normality of errors

One of the most basic assumptions we made for our regression model is the assumption of normality of the error terms. A lot of important results in linear regression analysis follows from this assumption. Although minor departures from normality is not an issue (and is often expected in most cases), major departures do create problems in fitting and reliability of the estimates. Thus, it is of utmost importance to verify this assumption. Following are a couple of ways to check this assumption :

- **Box-plots** or **histograms** of residuals (or standardized residuals) convey important information about the shape of the error distribution and the presence of outliers. Figure 3.3. shows these plots for the age-muscle mass data.



(a) Histogram (b) Boxplot

FIGURE 3.3: Histogram and Boxplots of standardized residuals for the age-muscle mass data.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

The plots seem to indicate a slight right skewness but it doesn't seem to serious. Moreover, two sided tests and confidence intervals are robust to violations of the normality assumption. So, the conclusions we have drawn before are still valid.

- A popular tool of assessing normality of the error distribution is to construct **Normal probability plots** of the residuals. Here, each residual is plotted against its expected value under normality. A linear plot suggested normality whereas a plot that deviates substantially from linearity suggests that the normality assumption may not be valid. Figure 3.4. shows the normal probability plot for the residuals of the age-muscle mass data.

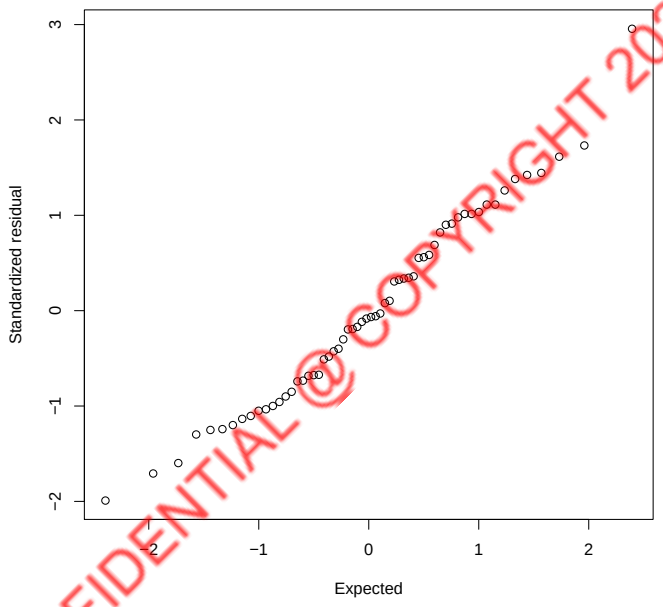


FIGURE 3.4: Normal probability plot for age-muscle mass data.

Since the pattern is pretty linear, the error distribution can be assumed to be normal. However, if the normality assumption is not satisfied, it is immediately reflected in the normal probability plots i.e. those tend to be nonlinear.

Note 1. *Residuals can also help us in detecting outliers. One rule of thumb is that, if the absolute value of the standardized residual for an observation is more than 3, that observation can be assumed to be an outlier.*

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

3.4 R codes

Following R codes were used to perform the analysis in this chapter. Explanations are provided alongside the codes whenever necessary.

1. Setting up the working directory:
   `setwd("D:/Desktop/SDA")`
   `getwd()`.

2. Importing the `Muscle-mass` dataset:
   `musclemass<-read.table("D:/Desktop/SDA/musclemass.txt",header=T)`
   `attach(musclemass)`.

3. Fitting least squares regression model to the `Muscle mass` data:
   `fit.mm<-lm(Y~X)`
   `summary(fit.mm)`.

4. Obtaining the standardized residuals:
   `std.resid<-rstandard(fit.mm)`.

5. Residuals vs Age, Fig 12.1:
   `plot(X,std.resid,xlab="Age",ylab="Standardized Residuals",pch=19)`
   `lines(c(40,90),c(0,0))`.

6. Residual plots for `Fossil` data, Fig 12.2:
   ```
   fossil<-read.table("D:/Desktop/SDA/fossil.txt",header=T)
   attach(fossil)
   fit.fossil<-lm(strontium.ratio~age)
   summary(fit.fossil)
   std.resid.fossil<-rstandard(fit.fossil)
   plot(age,strontium.ratio,xlab="Age", ylab="Strontium.ratio",pch=19)
   abline(fit.fossil)
   plot(age,std.resid.fossil,xlab="Age",ylab="Standardized Residuals",pch=19)
   lines(c(80,130),c(0,0))
   ```

7. Histogram and box-plot of residuals, Fig 12.3 :
   `hist(std.resid.fossil,xlab="Standardized Residuals",main="")`
   `boxplot(std.resid.fossil,xlab="Standardized Residuals",main="")`.

8. Normal probability plot of residuals, Fig 12.4 :
   `qqnorm(std.resid.fossil,xlab="Expected",ylab="Residual",main="")`.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Chapter 4

Multiple Linear Regression

4.1 Introduction

So long we have used a single explanatory variable (X) in the linear regression model to predict the unknown value of the response (Y). However, in many real life applications, the response (or outcome) of a process can depend on more than one explanatory variables. In those situations, we should ideally take ALL the explanatory variables into account to estimate (or predict) the unknown value of the response. Failing to do so would evidently result in loss of information about the true variability of the response and hence the resulting regression model will not be accurate enough for practical purposes.

Example 1.

The crime rate (number of crimes per 1000 residents)(say Y) of a particular region can depend on a lot of factors like the percentage of residents who are well educated (say X_1), the level of urbanization (X_2), the average income of the residents (X_3) etc. Thus, in order to accurately predict the true crime rate of a region, we should take all these factors into account because each of these give us some information about the crime rate in that region.

4.2 General Form

Suppose we have p explanatory variables, X_1, X_2,...,X_p corresponding to the response variable Y. Then the (population) multiple regression model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i, \quad i = 1, 2, ..., n \tag{4.1}$$

where ϵ_i's have a normal distribution with mean 0 and constant standard deviation σ^2 while n is the number of subjects/units. This is just an extension of the simple linear regression set up to p predictors.

The (least squares) predicted regression model is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip}$$

where $\hat{\beta}_k$ is the least squares estimate of $\beta_k, (k = 1, ..., p)$ obtained by minimizing the sum of squares of residuals - this is similar to what was done for simple linear regression. Clearly $\hat{\beta}_0$ is the estimated y-intercept while $(\hat{\beta}_1, ..., \hat{\beta}_p)$ are the estimated slopes corresponding to the predictors $(X_1, X_2, ..., X_p)$.

4.2.1 Interpretation of Coefficients

The meaning and interpretation of the parameters of the multiple regression model have the same spirit as those for the simple linear regression one. However, in order to interpret the effect of a predictor, we have to control for the others. This is because, in a multiple (linear) regression set up, the association pattern between the response and a predictor is NOT affected by any other predictor. This is so because the predictors and the response are related in an *additive* manner.

Example 2.

We have data on crime rate (Y), percent with high school education (X_1), percentage of residents living in an urban environment (X_2) and median income (X_3) (in thousands of Dollars) for all the counties of Florida, USA. Software generates the following estimated multiple regression model

$$\hat{Y} = 59.715 - 0.467X_1 + 0.697X_2 - 0.383X_3$$

We can interpret the parameters as follows :

Effect of Education : Since the slope of education is , crime rate of a county is related to education controlling for and . Specifically, the predicted crime rate of a county decreases by for 1 percent increase in the education rate i.e a county will be safer if more of its residents are educated.

Effect of Urbanization : Controlling for and , crime rate of a county is related to urbanization (since the slope is). In fact, the predicted crime rate of a county increases by for 1 percent increase in the urbanization rate i.e more urbanized the county, safer it is.

Effect of Income : Since the slope of income is , controlling for and , the median income of a county is related to its crime rate - for 1 thousand Dollar increase in the median income of the residents, the crime rate by i.e wealthier the residents of a county, safer it is.

Moreover, for the above regression model, the effect/slope of any predictor will remain the same for any value of the other predictors. For example, the slope of education would remain -0.467, no matter what values we assume for urbanization and income.

Note 2. *A basic difference between multiple and simple regression models is that for the former, in order to interpret the effect of a predictor, we* ***fix (or control for)*** *the other predictors but for*

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

the latter, we ***ignore*** *any other possible predictors in order to interpret the effect of a particular predictor.*

In the above example, suppose we regress Y only on X_2 (Urbanization). The regression equation is given by

$$\hat{Y} = 24.54 + 0.562X_2$$

Here education and income have been altogether ignored, NOT controlled. The slope of X_2 has also changed (decreased from 0.697 to 0.562). Thus, ignoring and controlling for a variable have different impact on the regression model.

4.3 Inferential Procedures

Now we will discuss some inferential procedures that can be performed on multiple regression models.

4.3.1 Analysis of Variance

As we have learnt in Chapter 9, ANOVA is a statistical procedure of decomposing the variability inherent in a dataset into different sources. In the context of regression analysis this will enable us to partition the total variability in the response (Y) into various components. Like other ANOVA formulations, regression ANOVA comes with sums of squares (SS), degrees of freedom (df), and mean squares (MS) as shown below.

1. Sums of Squares

ANOVA for regression involves the following sums of squares :

- The **Total (corrected) sum of squares** (SS_{Tot}) measures the total variability of the Y values around their overall mean. Its formula is

$$SS_{\text{Tot}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

- The **Regression sum of squares** (SS_{Regr}) measures the variability due to the regression equation. This is a function of the deviation of the predicted value, $\hat{Y}_i$, for each observation around the mean, $\bar{Y}$ as given below

$$SS_{\text{Regr}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2.$$

- The **Residual sum of squares** (SS_{Res}) is a function of the deviation of the observed values, Y_i around the corresponding predicted values $\hat{Y}_i$ given by

$$SS_{\text{Res}} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

It turns out that $SS_{\text{Tot}} = SS_{\text{Regr}} + SS_{\text{Res}}$.

2. Degrees of freedom

Each sum of squares has a degrees of freedom associated with it as follows :

$$df_{\text{Tot}} = n - 1, \qquad df_{\text{Regr}} = p, \qquad df_{\text{Res}} = n - p - 1.$$

Notice that $df_{\text{Tot}} = df_{\text{Regr}} + df_{\text{Res}}$. It is important to note that

- $df_{\text{Tot}} = n - 1$ because we are correcting for the actual mean.
- $df_{\text{Regr}} = p$ because there are p explanatory variables.
- $df_{\text{Res}} = n - p - 1$ because there are $p + 1$ estimable parameters $(\beta_0, \beta_1, ..., \beta_p)$.

3. Mean Squares

The mean squares are the sums of squares divided by the corresponding degrees of freedom i.e

$$MS_{\text{Regr}} = \frac{SS_{\text{Regr}}}{df_{\text{Regr}}}, \qquad MS_{\text{Res}} = \frac{SS_{\text{Res}}}{df_{\text{Res}}}.$$

4. ANOVA Table

The ANOVA table is shown below:

Source	df	SS	MS
Regression	df_{Regr}	SS_{Regr}	MS_{Regr}
Residual	df_{Res}	SS_{Res}	MS_{Res}
Total	df_{Tot}	SS_{Tot}	

FIGURE 4.1: Generic regression ANOVA table.

Now let us see how ANOVA can be applied to perform a significance test for linear association.

4.3.2 F test for β

Hypotheses

The F test is used to verify whether any of the predictors have a significant influence on the response. Hence, the null and alternative hypotheses will be

$$H_0 \quad : \quad \beta_1 = \beta_2 = ... = \beta_p = 0.$$
$$H_a \quad :$$

This test is based on the usual linear regression assumptions.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Test statistic

The F test statistic is given by

$$F = \frac{MS_{Regr}}{MS_{Res}} \tag{4.2}$$

A small F value would indicate higher evidence for H_0 while larger values would support H_a. However, this test is an inherently two-sided test i.e it can only indicate whether any of the predictors have a significant linear association with the response but cannot specify the direction of association. There is no such thing as a one-sided regression F test.

P-value

The above test statistic follows an F distribution with $df_1 = p$ and $df_2 = n-p-1$. Moreover, since the F statistic can only be positive, the p-value will be the tailed area above the observed F statistic value under a $F(p, n-p-1)$ distribution.

Decision

- If the p-value $\leq \alpha$, we H_0 and conclude that
- If p-value $> \alpha$, we fail to reject H_0 and conclude that

4.3.3 F test for Crime data

The ANOVA table for the `Crime rate` dataset is given below (*complete it*).

Source	Df	Sum of Squares	Mean Squares
Regression		24804.493	
Residual			
Total		52462.119	

TABLE 4.1: ANOVA table for age-muscle mass data.

The null and alternative hypotheses are :

$$H_0 \quad :$$
$$H_a \quad :$$

The F statistic is given by

with degrees of freedom . Since 18.834 is quite high a value, the p - value is close to 0. Thus we reject H_0 and conclude that at least one of education, urbanization or income has an effect on crime rate of a county.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Note 3. *To reiterate, it is good to perform the F test first to get an idea whether any of the explanatory variables has any influence on Y. If the p-value of the F test is large, there is no need to perform the individual t tests. But, if the p-value is small, we can then perform separate t tests for the individual effects to hunt down the predictor(s) which have a significant effect on Y.*

Figure 4.2 depicts a flow chart for the regression F test.

Large F value	Small F value
(much larger than 1)	(around 1 or less than 1)
↓	↓
Small p-value	Large p-value
↓	↓
Evidence against H_0 (for H_a)	No evidence against H_0 (for H_a)
↓	↓
Reject H_0	Fail to reject H_0
↓	↓
Evidence that Y depends on X	No evidence that Y depends on X

FIGURE 4.2: Results and interpretations of a regression F test.

4.3.4 T-tests of Regression Coefficients

As in the simple linear regression set up, we can test for the significance of the slope parameters for the multiple regression model (4.1). The necessary assumptions are the same as that for a simple linear regression model i.e

- Data is obtained through random sampling.
- The observations $(Y_i, X_{i1}, ..., X_{ip})$ are independent of each other.
- Linearity of the population regression model.
- Normality of the errors (vis-a-vis the response) for any given value of the predictors.
- Errors (vis-a-vis the response) has the same standard deviation for any given value of the predictors.

For testing the significance of a particular predictor, say X_k, the null and alternative hypotheses are given by

$$H_0 : \beta_k = 0 \quad vs \quad H_a : \beta_k \neq (\text{or} > \text{or} <)0$$

The above hypotheses tests for the dependence between X_k and Y controlling for the other predictors. The test statistic and corresponding p-values will given by any statistical software. However, the

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

degrees of freedom of the test statistic is $n-p-1$ since there are $p+1$ parameters in the multiple regression model.
As before, the decision rule will be

$$\text{p-value} < \alpha \Rightarrow \quad H_0.$$
$$\text{p-value} \geq \alpha \Rightarrow \quad H_0.$$

where the p-values are obtained in the usual manner.

Example 3.

For the Florida crime dataset, the above assumptions seem to be tentatively satisfied. Let us test whether income (X_3) has any effect in predicting crime rate controlling for urbanization (X_2) and education (X_1). Thus our hypotheses will be

The estimates and standard errors for the various predictors are given in the following table.

Predictor	Estimate	Standard Error
Intercept	59.715	28.59
Education	-0.467	0.554
Urbanization	0.697	0.129
Income	-0.383	0.941

Thus, the test statistic will be

Since there are 67 counties, the degrees of freedom of the above statistic will be
P-value : For the above value of the test statistic and degrees of freedom, the p-value can be shown to be much higher than 0.2.

Conclusion : At significance level of 0.05, we will H_0 since our p value is 0.05. Thus, we conclude that there is little evidence of any association between crime rate and income controlling for urbanization and education i.e income information of residents does not add significantly to our knowledge of crime rate if we already have information on urbanization and education rates of a county.

4.3.5 Confidence Intervals of Regression Coefficients

A 95% confidence interval of β_k is given by

$$\hat{\beta}_k \pm t_{\alpha/2,n-p-1} se(\hat{\beta}_k)$$

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

Example 4.

From the t-table, we have $t_{0.025,63} = 2.0$. Thus the 95% confidence interval of β_3 will be

Since the confidence interval contains 0, we are 95% confident that β_3 is not significantly different from 0 i.e controlling for urbanization and education, income doesn't seem to influence the crime rate of a county. Thus, both the two sided significance test and the confidence interval gives us the same conclusion about the slope.

Note 4. *Two sided tests and confidence intervals are robust to the violation of the normality assumption for large sample sizes ($n > 30$).*

4.4 Predictive Ability

Analogous to simple linear regression, predictive ability of a multiple linear regression model is measured using the *Coefficient of Multiple Determination*. This coefficient measures the proportion of variation in Y that is simultaneously explained by the set of predictors $(X_1, ..., X_p)$. As in the simple linear regression set up, R^2 ranges from 0 to 1 with higher values of R^2 indicating a better fitting model and vice versa.

4.4.1 R^2 through ANOVA

For simple linear regression (Chapter 10), we calculated R^2 by squaring the usual correlation coefficient (r). However, now that we know ANOVA, we can use those tools to calculate and understand the meaning of ANOVA specially in the context of multiple linear regression.

R^2 can be expressed as a ratio of SS_{Regr} and SS_{Tot} i.e

$$R^2 = \frac{SS_{\text{Regr}}}{SS_{\text{Tot}}}.$$

where

$$SS_{\text{Tot}} = SS_{\text{Regr}} + SS_{\text{Res}},$$

Thus, R^2 is the proportion of SS_{Tot} that is accounted for by

- If our regression equation has good predictive ability, SS_{Res} is and hence most of SS_{Tot} comes from , resulting in a value of R^2 i.e R^2 is close to
- If our regression equation has poor predictive ability, SS_{Res} is and hence most of SS_{Tot} comes from instead of SS_{Regr}, resulting in a value of R^2 i.e R^2 is close to

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

For the `Crime` data,

$$R^2 = \frac{SS_{\text{Regr}}}{SS_{\text{Tot}}} =$$

Interpretation :

4.4.2 Adjusted R^2 & Multicollinearity

R^2 can only increase when additional predictor variables are added to the model. However, increasing the predictors will also increase the number of parameters and hence the computational cost. In order to achieve a trade off between these two factors, an adjusted coefficient of multiple determination has been proposed, given by

$$R_a^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1-R^2)$$

In fact, R_a^2 can even decrease with the addition of a predictor variable in the regression model if the new predictor does not result in a significant improvement of the model fit.

As we have seen, for the crime data, R^2 for the full model is .473 i.e Education, Urbanization and Income taken together explains about of the total variation in crime rate. Now, let us test the amount by which R^2 increase as we include more and more predictors. We start by only including Urbanization (since it has the highest correlation with crime rate) and add on the other predictors. The following table shows the R^2 values for each case.

Predictors	U	(U, I)	(U, E)	(U, I, E)
R^2	0.459	0.467	0.471	0.473

It is clear that once we have included urbanization, income and education does not add a significant amount of information on the variability of crime rate. This is because all the predictors are highly correlated and hence have a high degree of overlap of the information they possess on crime rate (r(I, E) = .793, r(U, I) = 0.731, r(U, E) = 0.791). As a result, once one of the predictors, say urbanization is included in the model, the rest of the predictors does not bring in significant additional information on crime rate and thus becomes redundant. The following table shows the values of adjusted R^2 corresponding to various combinations of predictors.

Predictors	U	(U, I)	(U, E)	(U, I, E)
R_a^2	.4507	.4503	.4545	.4479

Clearly, R_a^2 identifies the model with urbanization and education as a "better" model than the one with all the predictors. In fact, it seems that the optimal model is the one with

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

The above phenomenon is known as **Multicollinearity** and it affects the manner in which predictors relate to themselves and to the response as a whole - hence it must be dealt with care. Last but not the least, it must be remembered that in a multiple regression setting, a regression coefficient only reflects the *partial* effect of the corresponding predictor on the response conditional on the predictors, and not the absolute one.

4.4.3 Measuring Multicollinearity

Multicollinearity is measured using what is known as the **Variance Inflation Factor** or VIF. The more severe the multicollinearity, higher is the VIF. We will treat 3 as our threshold (although there are universally accepted thresholds) i.e if the VIF of a predictor is 3 or above, it can be deemed redundant (given the rest) and hence can be dropped. The model should then be refitted with the rest of the predictors and this exercise should be continued until all the predictors have $VIF < 3$.

Example 5. *For the Florida crime data set, we have the following* `R` *output*

Predictor	Estimate	Standard Error	T	P-value	VIF
Intercept	59.715	28.59	2.09	0.041	
Education	-0.467	0.554	-.84	0.403	**3.627**
Urbanization	0.697	0.129	5.40	0.00	**2.893**
Income	-0.383	0.941	-0.41	0.685	**2.916**

Based on the VIF values, we conclude that can be dropped. Once this is done, the model should be refitted with and and the VIFs checked again. In fact, and can be dropped as well since they have high p-values. It should be remembered that dropping a variable from a model based on VIF *does not* mean that the variable has no effect on the response. It just means that its effect is adequately explained by the remaining variables.

4.5 Regression Diagnostics

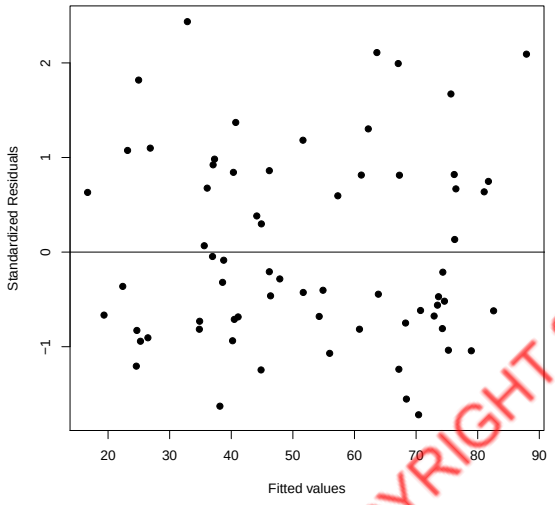
As for simple linear regression, we can use residual analysis to verify the assumptions of the multiple regression model, specifically those relating to normality, linearity and constant variance. Thus, this is a nice tool to test for the overall appropriateness of the model. For example,

- Plot of residuals (or standardized residuals) against fitted values can help us to test for linearity of the regression model and the constancy of error variances.
- Plot of the residuals against each of the predictors can be used to check whether the response is linearly related to those predictors controlling for the others.
- Boxplots/histograms and normal probability plots of the residuals can be used to check for the validity of the normal distributional assumption.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

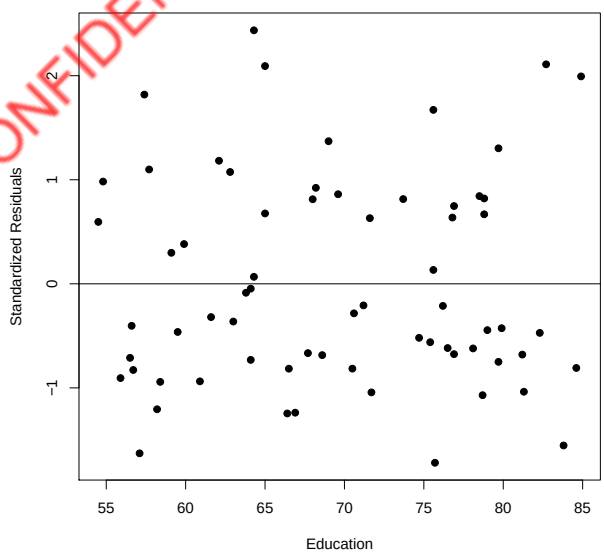
Example 6. *For the crime data set, the residual analysis is given below.*

1. The plot of the standardized residuals against the fitted values ($\hat{Y}$) are shown below



   Since there is not definite pattern, we can conclude that the linearity and constant variance assumptions are valid.

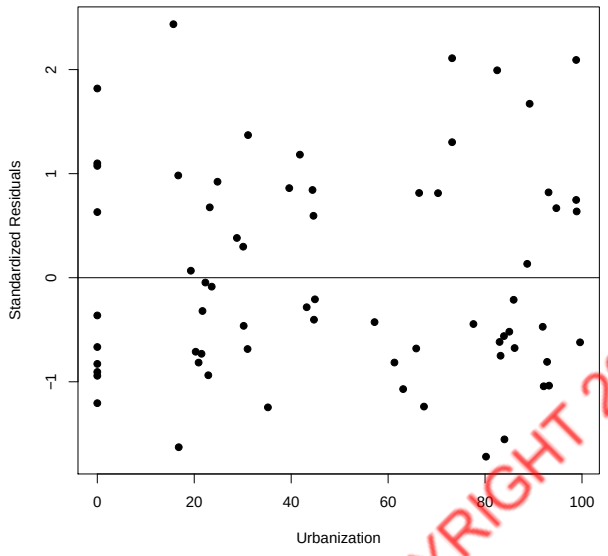
2. The residual plot corresponding to education is shown below



   The residuals fluctuate more or less randomly about 0 with no noticeable trend or variation. Thus we conclude that crime rate can be assumed to be linearly related to education.

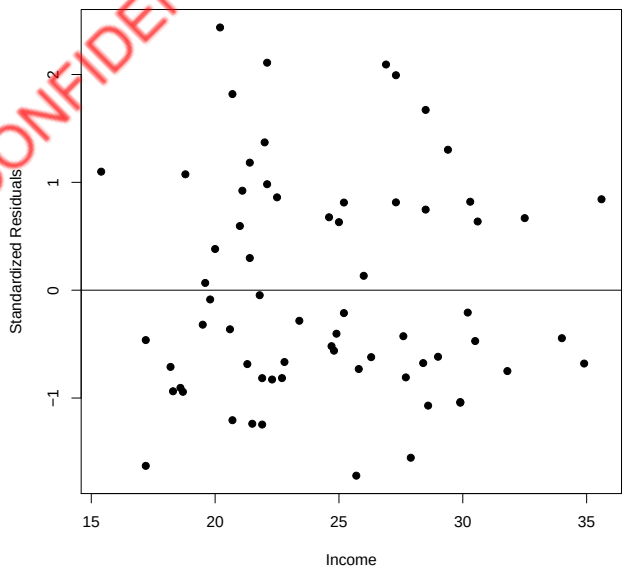
DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

3. The residual plot corresponding to urbanization is shown below.



Here also the residuals fluctuate more or less randomly about 0 with no noticeable trend or variation. Thus crime rate can be assumed to be linearly related to urbanization.

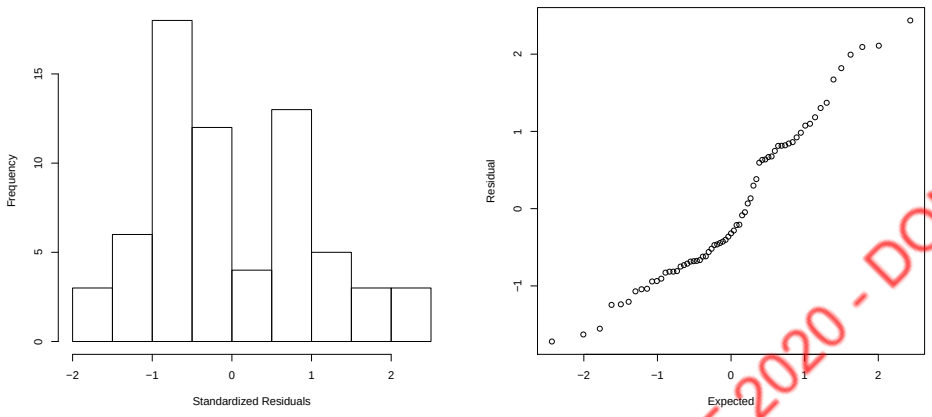
4. The following figure shows the residual plot against income.



Here also the residuals seem to have a random pattern about the 0-line. Hence, we can assume that crime rate is linearly related to income.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

5. The histogram and normal probability plots of the standardized residuals are shown below



(a) Histogram (b) Normal Probability Plot

The above plots indicate that the normal distributional assumptions (for the error/response terms) may have been violated. However, two sided tests and confidence intervals of the slope parameters are robust against violation of this assumption. Thus, the conclusions we have drawn earlier regarding the effects of the explanatory variables on the crime rate still holds.

DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE

4.6 R codes

Following R codes were used to perform the analysis in this chapter. Explanations are provided alongside the codes whenever necessary.

1. Importing the `Crime rate` dataset :
   ```
   flcrime<-read.csv("D:/Desktop/SDA/flcrime.csv",header=TRUE)
   crime<-flcrime$crime
   edu<-flcrime$education
   urb<-flcrime$urbanization
   inc<-flcrime$income.
   ```

2. Fitting least squares regression model to the `Crime` data and obtaining parameter estimates, test statistics, p-values, R^2 etc :
   ```
   fit.crime<-lm(crime~edu+urb+inc)
   summary(fit.crime).
   ```

3. ANOVA and F test for Crime data:
   ```
   anova(fit.crime).
   ```

4. Obtaining the standardized residuals:
   ```
   std.resid<-rstandard(fit.crime).
   ```

5. Residuals vs Fitted values :
   ```
   plot(fitted(fit.crime),std.resid,xlab="Education",ylab="Standardized Residuals",pch=19)
   lines(c(50,90),c(0,0)).
   ```

6. Residual plot for education :
   ```
   plot(edu,std.resid,xlab="Education",ylab="Standardized Residuals",pch=19)
   lines(c(50,90),c(0,0)).
   ```

7. Residual plot for urbanisation :
   ```
   plot(urb,std.resid,xlab="Urbanization",ylab="Standardized Residuals",pch=19)
   lines(c(-10,110),c(0,0)).
   ```

8. Residual plot for income :
   ```
   plot(inc,std.resid,xlab="Income",ylab="Standardized Residuals",pch=19)
   lines(c(10,40),c(0,0)).
   ```

9. Boxplot, Histogram and Normal probability plot of residuals :
   ```
   hist(std.resid,xlab="Standardized Residuals",main="")
   boxplot(std.resid,xlab="Standardized Residuals",main="")
   qqnorm(std.resid,xlab="Expected",ylab="Residual",main="").
   ```


DINESH BABU - CONFIDENTIAL @ COPYRIGHT 2020 - DON'T DISTRIBUTE