Learning in R

Optimization & Linear Regression

Dr. Abhishek Singh

abhishek@nith.ac.in

Department of Mathematics, NIT Hamirpur, (H.P.)

July 8, 2019

Outline

Learning Objectives

- Define different types of optimization techniques used in linear regression analysis.
- Define least squares method to obtain regression coefficients and how they are assessed.
- Define linear regression to establish the cause and effect relationship between variables.
- Describe the kinds of evidence that would be relevant to use linear regression model.
- Hands on practice in R software.

Introduction

- Statistical science is concerned with optimal decision making under uncertainty in various contexts.
  1. Collection of data
  2. Analysis
  3. Interpretation of available data
- How do researchers make optimal decisions?

Answer

- Optimization!!!
- It is the act of achieving the best possible result under given circumstances.
- In linear regression extensive use is made of optimization techniques such as:
  1. **least squares**
  2. maximum likelihood estimation
  3. most powerful tests

Regression Analysis

Regression analysis is a mathematical measure of the average relationship between tow or more variables in terms of the original units of the data.

- In regression analysis there are two types of variables.

1. **Dependent Variable:** The variable whose value is influenced or is to be predicted is called dependent variable.
2. **Independent Variable:** The variable which influences the values or is used for prediction is called independent variable.

Regression Analysis

- If the variables in a bivariate distribution are related:
  - The points in the scatter diagram will cluster round some curve called the *curve of regression.*
  - If the curve is a straight line, it is called the line of regression and there is said to be linear regression between the variables.
  - Otherwise regression is said to be cuvilinear.

Line of Regression

- The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable.
- Thus the line of regression is the line of *best fit* and is obtained by the principle of least squares.

Principle of Least Squares

- Let us suppose that in the bivariate distribution $(x_i, y_i); i = 1, 2, ..., n$; Y is dependent variable and X is independent variable.
- Let the line of regression of Y on X be:

$$Y = a + bx \tag{1}$$

- The above equation represents a family of straight lines for different values of the arbitrary constants 'a' and 'b' so that the above line is the line of best fit.

Principle of Least Squares

- Legendre's principle of least squares consists in minimising the sum of the squares of the deviations of the actual values of y from their estimated values as given by the line of best fit.
- The term best fit is interpreted in accordance with principle of least squares.

Principle of Least Squares

Let $P_i\ (x_i, y_i)$ be any general point in the scatter diagram. Draw $P_iM\ \perp$ to x-axis meeting the point H_i. Abscissa of H_i is x_i and since H_i lies on the line, its ordinate is $a + bx_i$. Hence the co-ordinates of H_i are

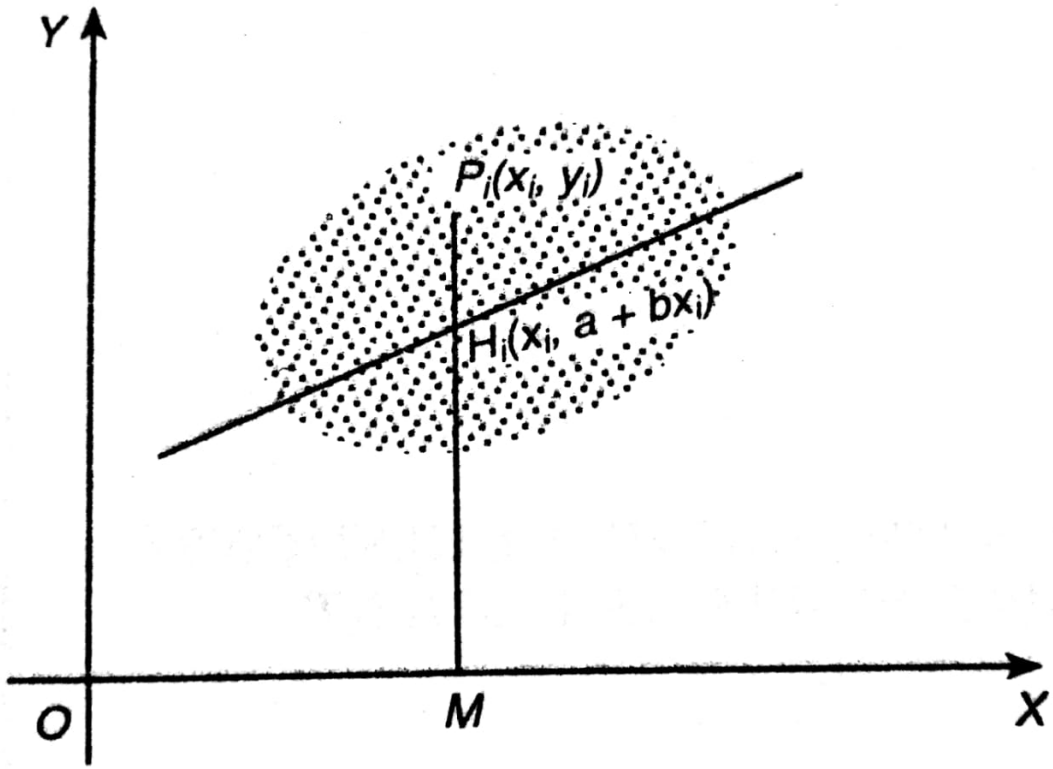


Figure: Best Fit

Principle of Least Squares

- According to the principle of least squares, we have to determine a and b so that

$$E = \sum_{i=1}^{n} P_i H_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

  is minimum.

- From the principle of maxima and minima, the partial derivative of E, with respect to a and b should vanish separately, i.e.,

Principle of Least Squares

$$\frac{\delta E}{\delta a} = 0 = -2\sum_{i=1}^{n}(y_i - a - bx_i)$$

$$\Rightarrow \quad \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \tag{2}$$

$$\frac{\delta E}{\delta b} = 0 = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i)$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \tag{3}$$

Principle of Least Squares

- Equations (2) and (3) are known as the **normal equations** for estimating a and b.
- All the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i$ *and* $\sum_{i=1}^{n} x_i y_i$, can be obtained from the given set of points $(x_i, y_i); i = 1, 2, ..., n$.
- Now, equations (2) and (3) can be solved for a and b.
- With the values of a and b so obtained, equation (1) is the *line of best fit* to the given set of points (x_i, y_i); $i = 1, 2, ..., n$.

Simple Linear Regression

- A model with a single regressor x that has a relationship with a response y that is a straight line.
- This simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Where the intercept β_0 and the slope β_1 are unknown constants.
- ϵ is a random error component.

Random Error/Residual

- The difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$ is a **residual.**
- Mathematically the i^{th} residual is:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \ i = 1, 2, ..., n$$

- Residuals play an important role in investigating **model adequacy** and in detecting departures from the underlying assumptions.

Assumptions

1. **Zero mean value of error e_i:** The residuals are assumed to have **mean zero,** i.e., $E(e_i) = 0$
   - This assumption implies that there is no **specification bias** or **specification error** in the model used in empirical analysis.
2. **Homoscedasticity:** equal (homo) & spread (scedasticity). It is assumed that **variance of residuals are constant,** i.e., $var(e_i) = \sigma^2$
3. **No autocorrelation between the errors:** It is also assumed that **errors are uncorrelated,** i.e., $cov(u_i, u_j) = 0$

Regression Coefficients Interpretation

- The parameters β_0 and β_1 are known as **regression coefficients.**
  - The slope β_1 is the change in the mean of the distribution of y produced by a unit change in x.
  - If the range of data on x includes $x = 0$, then the intercept β_0 is the mean of the distribution of the response y when $x = 0$.
  - If the range of x does not include zero, then β_0 has no practical interpretation.

Least Squares Estimation of β_0 & β_1

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$where,\ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$and,\ S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

The Rocket Propellant Data

- We have **twenty observations** on shear strength and the age of the corresponding batch of propellant are shown in the following figure:

Observation, i	Shear Strength, y_i (psi)	Age of Propellant, x_i (weeks)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

Figure: Rocket Propellant Data

The Rocket Propellant Data

- The linear regression model **(least squares fit)** is:

$$\hat{y} = 2627.82 - 37.15x$$

Interpretation of Regression Coefficients:

- The slope -37.15 is interpreted as the **average weekly decrease** in propellant shear strength due to the age of the propellant.
- Since the lower limit of the $x's$ is near the origin, the intercept 2627.82 represents the shear strength in a batch of propellant immediately following manufacture.

Diagnostic Checks

- After obtaining the least squares fit, a number of interesting questions come to mind:
  1. How well does this equation fit the data?
  2. Is the model likely to be useful as a predictor?
  3. Are any of the basic assumptions violated, and if so, how serious is this?
- All of these issues must be investigated before the model is fi nally adopted for use.

Summary

- Optimization techniques are extensively used in regression analysis.
- The least squares method of optimization is used when the form of the distribution of errors is not known.
- Least squares methods provides normal equations to find out the value of regression coefficients.
- It is one of the most widely used method of estimation in regression analysis.

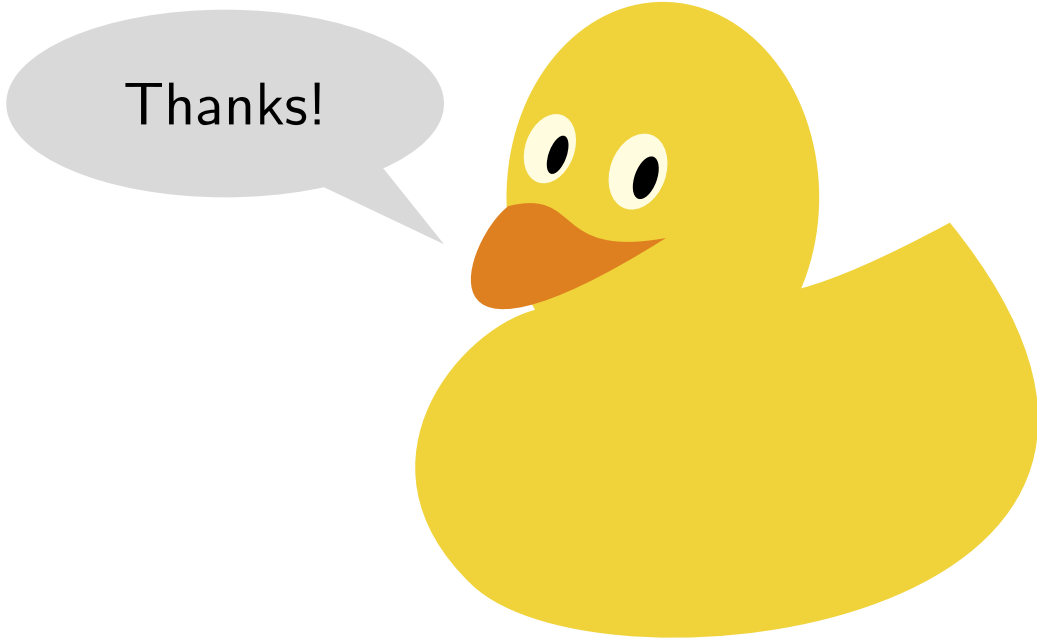
Summary

- Linear regression model establish the cause & effect relationship between independent and dependent variable.
- The random error term e_i should have zero mean and constant variance and follow normal distribution.
- The slope β_1 is the change in the mean of the distribution of y produced by a unit change in x.
- If the range of data on x includes $x = 0$, then the intercept β_0 is the mean of the distribution of the response y when $x = 0$.
- If the range of x does not include zero, then β_0 has no practical interpretation.

For Further Reading I

S.C. Gupta, and V.K. Kapoor.
Fundamentals of Mathematical Statistics.
Sultan Chand & Sons, 2012.

D.C. Montgomery, E.A. Peck, and G.G. Vinning.
Introduction to Linear Regression Analysis.
Wiley, 2003.


Thanks!