

Least Squares And Regression Analysis

Hossein Falaki
Computer Science Department,
University of California, Los Angeles
falaki@cs.ucla.edu

1 Introduction

This note is an introduction to *regression analysis*. We also introduce *Least Squares* and prove the *Gauss-Markov* theorem on the optimality of the least square method.

2 History of Regression

A simple form of regression was used by Gauss to improve Euler's equations for Jupiter's orbit. Jupiter's motion is affected by both the Sun and Saturn. Euler derived Jupiter's location as:

$$\begin{aligned}\varphi = & \eta + 23525'' \sin q + 168'' \sin 2q + 32'' \sin 2w - 257'' \sin(w - q) \\ & - 243'' \sin(2w - p) + m'' - x'' \sin q + y'' \sin 2q + q'' \sin(w - p) \\ & - u(\alpha + 365v + p) \cos(w - p) + Nn'' - 11405R'' \cos p + (1/600)k'' \cos 2p\end{aligned}$$

Where η is the location of Jupiter based on two-body equations (affected by the Sun's gravity). In this equation $(\varphi, \eta, q, w, p, N, v)$ are given by observations and $(x, y, m, z, \alpha, k, n, u)$ are fixed unknown constants. Euler had 75 observations. He used seven of them to find the value of unknowns. It is important to note that observations are not perfectly accurate and each one comes with an *observation error*.

In this example the equation system consists of seven unknowns but 75 equations. The critical difference is that each observation comes with an error, therefore it makes sense to use all the equations. Laplace improved Euler's result by combining all the 75 observations into seven equations:

$$\left\{ \begin{array}{l} 1 : eqn1 + \dots + eqn75 \\ 2 : eqn1 - \dots + eqn75 \\ \vdots \end{array} \right.$$

observations	Y	regressors(X)			
1	y_1	x_{11}	x_{12}	\dots	x_{1p}
2	y_2	x_{21}	x_{22}	\dots	x_{2p}
\vdots	\vdots				\vdots
n	y_n	x_{n1}	x_{n2}	\dots	x_{np}

Table 1: Assumption in a regression analysis

3 Problem Statement

Assume that we know a priori that an observation is linear to a set of known variables (regressors)¹. For each observation:

$$y_i \doteq x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + x_{ip}\beta_p \quad (1)$$

To write the n equations in compact matrix form, assume that Y and each X_i are vectors:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}$$

We can represent the regressors as the following $n \times p$ matrix:

$$X = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix}$$

In this notation all the equations can be written in the following compact form:

$$Y_{n \times 1} \doteq X_{n \times p} \beta_{p \times 1} \quad (2)$$

where β is the vector of all the β_i values.

Using this notation a general form of Laplace's combination of equations can be expressed with a matrix of weights. In other words, when there are p regressors, we need p sets of weights each consisting of n values. Each set will result in an equation, and the p equations can be used to solve the unknowns.

$$\sum_{i=1}^n w_{ij} y_i \doteq \sum_{i=1}^n w_{ij} X_i^T \beta \quad j = 1, \dots, p$$

¹This assumption is rarely true.

If we represent all w_{ij} in matrix form:

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2p} \\ \vdots & \vdots & & \vdots \\ w_{i1} & w_{i2} & \dots & w_{ip} \end{pmatrix}$$

The equations can be summarized as:

$$W_{p \times n}^T Y_{n \times 1} \doteq W_{p \times n}^T X_{n \times p} \beta_{p \times 1} \quad (3)$$

In Equation 3 the dimension of the matrices are stated for clarity.

An obvious question to ask is the *optimal value* of W . The *Gauss-Markov* theorem states that the optimal value of W is X . Before proving this result, we need to introduce the *least square method* and define what optimality for W means in this context. In the next section we define the optimality criteria of W in the simple one dimensional case.

4 One Dimensional Case

In the special case of $p = 1$, the linearity assumption translates to:

$$y_i = x_i \beta_{true} + \epsilon_i \quad i = 1, \dots, n \quad (4)$$

In this statistical formulation β_{true} is the true value that we are trying to discover, and ϵ_i , are observation errors. They are assumed to be uncorrelated, zero-mean with constant variance. i.e.,

$$\begin{aligned} E(\epsilon_i) &= 0 \\ Var(\epsilon_i) &= \sigma^2 \end{aligned}$$

The linear combination of equations would be:

$$\sum_{i=1}^n w_i y_i \doteq \sum_{i=1}^n w_i x_i \beta$$

Therefore:

$$\hat{\beta} = \frac{\langle W, Y \rangle}{\langle W, X \rangle}$$

For each W , under hypothetical repeated sampling we get a different set of ϵ_i s and therefore a different Y , which leads to a different answer for $\hat{\beta}$. In this case, $E(\hat{\beta})$ is the long-run average, and $Var(\hat{\beta})$ is the long-run variability.

Replacing the true value of Y in our solution for $\hat{\beta}$ gives:

$$\begin{aligned} \hat{\beta} &= \frac{\langle W, X \beta_{true} + \epsilon \rangle}{\langle W, X \rangle} \\ &= \beta_{true} + \frac{\langle W, \epsilon \rangle}{\langle W, X \rangle} \end{aligned}$$

Lets derive $E(\hat{\beta})$ and $Var(\hat{\beta})$:

$$\begin{aligned}
E(\hat{\beta}) &= E(\beta_{true} + \frac{\langle W, \epsilon \rangle}{W, X}) \\
&= \beta_{true} + \frac{E(\langle W, \epsilon \rangle)}{\langle W, X \rangle} \\
&= \beta_{true} + 0 \\
&= \beta_{true}
\end{aligned}$$

Therefore, under repetition $\hat{\beta}$ values fluctuate around the true value of β .

$$\begin{aligned}
Var(\hat{\beta}) &= Var(\beta_{true} + \frac{\langle W, \epsilon \rangle}{W, X}) \\
&= \frac{E(\langle W, \epsilon \rangle^2)}{\langle W, X \rangle^2} = \frac{Var(\sum^n w_i \epsilon_i)}{\langle W, X \rangle^2} \\
&= \frac{\sum Var(w_i \epsilon_i)}{\langle W, X \rangle^2} = \frac{\sum w_i^2 \sigma^2}{\langle W, X \rangle^2} \\
&= \frac{\|W\|^2 \sigma^2}{\|W\|^2 \|X\|^2 Cos^2 \theta} = \frac{\sigma^2}{\|X\|^2 Cos^2 \theta}
\end{aligned}$$

This gives a criteria to define optimal W . The variance of $\hat{\beta}$ is minimized when $Cos\theta = \pm 1$. Therefore $W \propto X$. Such W is called the *Best Linear Unbiased Estimator* (BLUE). In other words, $W = X$ is an optimal value for W . This is the simple one-dimensional case of the Gauss-Markov theorem. In the next section we consider the general case.

4.1 General Case

Using the formulation of Section 3, we can easily derive the general solution to β .

$$W^T Y = W^T X \beta$$

For any given W , β is:

$$\hat{\beta} = (W^T X)^{-1} W^T Y \quad (5)$$

In the next section we introduce the least square method to find β and then we will prove in the general case that $W = X$ gives the optimal regression.

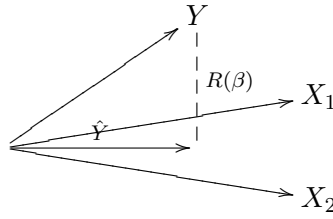
5 Least Square Method

Consider the assumptions presented in Table 1, and the linear model:

$$Y = X\beta + \epsilon \quad (6)$$

Where β is a vector of p values, and ϵ is a vector of n values.

The least square method for regression finds β such that the residuals are minimized.



The residual is defined as follows:

$$R(\beta) = \|Y - X\beta\|^2 = \sum_{i=1}^n (y_i - X_i^T \beta)^2 \quad (7)$$

Lets derive the β that minimizes $R(\beta)$.

$$\begin{aligned} \frac{\partial R}{\partial \beta_j} &= -2 \sum_{i=1}^n [y_i - (x_{i1}\beta_1 + \dots + x_{ip}\beta_p)] x_{ij} \quad j = 1, \dots, p \\ \Rightarrow \frac{\partial R}{\partial \beta_i} &= -2 \langle Y - X\beta, X_j \rangle = 0 \end{aligned}$$

To minimize the residual we must find β such that $\frac{\partial R}{\partial \beta_i}$ for all j .

$$X^T(Y - X\beta) = 0 \quad \Rightarrow \quad X^T = X^T X \beta$$

Therefore:

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T Y \quad (8)$$

The projected value of least square would be:

$$\hat{Y} = X\hat{\beta}_{LS} = \underbrace{X(X^T X)^{-1} X^T}_{H} Y \quad (9)$$

Lemma 5.1. Cauchy-Schwarz inequality: For any vector Y :

$$Y^T Y \geq Y^T H Y \quad (10)$$

Proof. It is obvious that the length of the projected vector, \hat{Y} is less or equal to the length of the original vector, Y . i.e.,

$$\|Y\|^2 \geq \|\hat{Y}\|^2$$

By expanding we get:

$$Y^T Y \geq (HY)^T (HY) = Y^T H Y$$

The last equation is because H is an operator that projects a vector in to the space defined by X_i s. Therefore projecting a vector multiple times gives the same result \square

Now we can prove the Gauss-Markov theorem.

Theorem 5.2. *The least square method gives the optimal β .*

In other words, β_{LS} is the same as optimal β among all the linear combinations: $\hat{\beta} = (W^T X)^{-1} W^T Y$.

Proof. It is obvious that setting $W = X$ gives the same result as the least square method, but we need to prove that this W is indeed optimal by the definition of optimality given in Section 4. Suppose the real model is:

$$y = x_1\beta_1 + x_2\beta_2 + \dots + x_p\beta_p$$

Lets call the vector of $\beta_1 \dots \beta_p$, β_{true} . Consider the general $\hat{\beta}$ with a given W :

$$\hat{\beta} = (W^T X)^{-1} W^T Y$$

Lets use this “learned” value to predict Y with a new set of explanatory variables, X_0 .

$$\hat{y}_0 = x_{01}\hat{\beta}_1 + x_{02}\hat{\beta}_2 + \dots + x_{0p}\hat{\beta}_p = X_0^T \hat{\beta}$$

whereas the true value of y is:

$$x_{01}\beta_1 + x_{02}\beta_2 + \dots + x_{0p}\beta_p = X_0^T \beta_{true}$$

Replacing $\hat{\beta}$ with its value from Equation 5 we get:

$$\begin{aligned} Y_0 &= X_0^T \hat{\beta} = X_0^T (W X)^{-1} W^T Y \\ &= X_0^T (W^T X)^{-1} W^T (X \beta_{true} + \epsilon) \\ &= X_0^T \beta_{true} + X_0^T (W^T X)^{-1} W^T \epsilon \end{aligned}$$

For simpler notation: $A^T = X_0^T (W^T X)^{-1} W^T$

We can now find the expectation and variance of Y_0 :

$$\begin{aligned} E(X_0^T \hat{\beta}) &= X_0 \beta_{true} \\ Var(X_0^T \hat{\beta}) &= \|A\|^2 \sigma^2 \end{aligned}$$

If $W = X$, then $\|A\|^2$ would be:

$$\begin{aligned} \|A_{LS}\|^2 &= A_{LS}^T A_{LS} \\ &= (X_0^T (X^T X)^{-1} X^T) (X (X^T X)^{-1} X_0) \\ &= X_0^T (X^T X)^{-1} X_0 \end{aligned}$$

From the Cauchy-Schwarz inequality we know that:

$$\|A\|^2 \geq A^T H A$$

The right hand side of this inequality is:

$$\begin{aligned} A^T H A &= A^T X (X^T X)^{-1} X^T A \\ &= (X_0^T (W^T X)^{-1} W^T) (X (X^T X)^{-1} X^T) (W (X^T W)^{-1} X_0) \\ &= X_0^T (X^T X)^{-1} X_0 \end{aligned}$$

The last term is equal to $\|A_{LS}\|^2$. This proves that the variance of $X_0 \hat{\beta}$ for any $W = X$ is the minimum. Therefore the least square solution is the optimal solution to β .

□

6 Acknowledgment

This note is based on Prof. Ying Nian Wu's lectures on Theoretical Statistics at UCLA.