M – 28

Least Squares Estimator for β in Multiple Regression Estimation

Tubagus Pamungkas
Pend Matematika Unrika Batam
Batam, Indonesia
tubagus@unrika.ac.id

***Abstract*—**Regression analysis has been developed to study the pattern and measure the statistical relationship between two or more variables . Mechanical analysis that attempts to explain the relations between two or more variables or more specifically the relationship between variables containing causation is called regression analysis . The procedure is based on the analysis of joint probability distribution variables. If this relationship can be expressed in a mathematical equation , it can be used in everyday purposes , for example to make predictions , fortune telling , and so on

Overall regression test with paramater (parametric regression) and regression without parameter (semiparametric regression) in advance will be discussed for parametric regression parameters $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, …, $\boldsymbol{\beta_{m-1}}$, $\boldsymbol{\beta_m}$ which is an element β in a model y = Xβ + ε. in this case will assume that y distribution Nn (Xβ, σ2I), where X notated n x (m+1) from rank m+1 < n.

Least Square approach to estimation of the β in the fixed model, for the parameter $\beta_0, \beta_1, \beta_2, \ldots \ldots \beta_m$, estimators that minimize the sum of squares of deviations of the n observed y from their predicted values $\hat{y}$.

Keywords: *regression, estimator, least square.*

I. INTRODUCTION

In everyday life there are things that can be solved using mathematics, statistics is one way to collect data, process, analyze and conclude. Regression analysis is a technique to look at the correlation between two or more variables and then estimation become a model that can be an equation that can connect the dependent variable of the independent variable. Many paper non-parametric regression estimation for efficiency in production on the independent variables in certain procedures to explain the factors that may affect the performance of the dependent variable. The regression model that handles these situations requires a set of equations (one single equation alone is not enough) that need to be solved simultaneously and this model is known as econometric models. In addition, conventional approaches to inference used in this paper is invalid because an elaborate serial correlation, unknown among the estimated efficiencies. Authors first describe a decent data for models like this.

The mathematical equations that allow to forecast the values of a dependent variable of one or more independent variables called the regression equation. The term is derived from the results of observations made Sir Francis Galton (1822 - 1911) that compares the height of a boy with his father's height. Galton states that the height of the boys from the father high on several generations later tended to "regressed" close to the average population.

Watson (1937) uses a regression of leaf area to estimate the average extent in a factory. The procedure is to weigh the whole leaves of the plant. For a small sample of the leaves, broad and weight of each leaf

has been set. On average regression on the weight of the leaves, the core of the application is that the weight of the leaves can be found quickly but the determination is time consuming.

II. DISCUSSION

The muptiple linear regression model can be expressed as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + \beta_m x_m + \varepsilon$ The$\beta_0, \beta_1, \beta_2, \ldots . \beta_m$ can be estimated by the least square approximation as long as the model is linear in the $\beta_0, \beta_1, \beta_2, \ldots . \beta_m$

We have n observation, with this equation

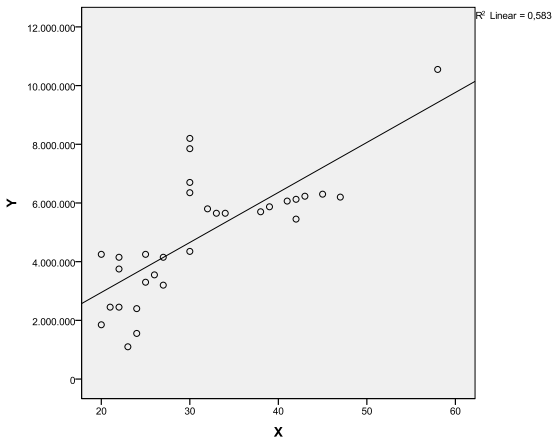
$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_m x_{1m} + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_m x_{2m} + \varepsilon_2$$
$$y_3 = \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \cdots + \beta_m x_{3m} + \varepsilon_3$$
$$\vdots$$
$$y_{n-1} = \beta_0 + \beta_1 x_{(n-1)1} + \beta_2 x_{(n-1)2} + \beta_3 x_{(n-1)3} + \cdots + \beta_m x_{(n-1)m} + \varepsilon_{(n-1)}$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_m x_{nm} + \varepsilon_n$$

We can show thats equation in matrix formula $y = X\beta + \varepsilon$

With $y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix}$ and than $X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ 1 & x_{31} & x_{32} & \cdots & x_{3m} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$

Estimator of β

For the all of parameter $\beta_0, \beta_1, \beta_2, \ldots . \beta_m$ and for minimize the sum of square of deviations in an observation we can predict value of $\hat{y}$ with $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots . , \hat{\beta}_m$ so we can use equation $y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_m x_{1m} + \varepsilon_1$ until $y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_m x_{nm} + \varepsilon_n$ with the equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_m x_{im}$



the linear line $\hat{y}$ shows the distribution of the dots indicate y, so it can be seen that $\hat{y} = \varepsilon + y$ in other words $\varepsilon = \hat{y} - y$ withdrawal procedure known regression line is the least squares method (ordinary least squares) or better known with the term OLS. This method choose a regression line to make the sum of squared vertical distances of the points through which the straight line as small as possible, whereby if the multiple linear regression model of the population is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + \beta_m x_m + \mathcal{E}$ where as the estimation model of multiple linear regression is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \cdots + \hat{\beta}_m x_{im}$ so meaning OLS estimates on multiple linear regression there are:

Population regression equation : $y = X\hat{\beta} + \varepsilon$

Residual (estimate of random error) : $\varepsilon = y - X\hat{\beta}$

Sum of Squares Error (SSE) : $\hat{\varepsilon}'\hat{\varepsilon} = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)$

$= y'y - 2\hat{\beta}'y + \hat{\beta}'X'X\hat{\beta}$

minimize SSE : $\frac{\partial(\hat{\varepsilon}'\hat{\varepsilon})}{\partial\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$

Estimator OLS : $\hat{\beta} = (X'X)^{-1}X'y$

Least square estimation for $\hat{\beta}$

From estimation OLS $\hat{\beta} = (X'X)^{-1}X'y$ if $E(y) = X\beta$ then $\hat{\beta}$ is an unbiased estimator for β so we can show that $E\left(\hat{\beta}\right) = E((X'X)^{-1}X'y)$

$E\left(\hat{\beta}\right) = (X'X)^{-1}X'E(y)$=β

III. CONCLUSION

Linear regression estimates were made to improve the accuracy by using additional variables that correlated with. When y_i the relationship between x_i and y_i tested, it was found that although the relationship may be a linear approach, the line is not via the point of origin. These results suggest an estimate based on a linear regression of the y_i from x_i better than the ratio of two variables.

Estimated regression is consistent, in simple terms when the sample consists of all units of the population $\bar{x} = \bar{X}$ and reduce the regression estimates $\bar{Y}$. As will be shown, in general regression estimates are biased, but the ratio of the bias is to the standard error becomes smaller when large samples.

With an appropriate choose of β , regression estimates included as special cases on average per unit and the estimated ratio, when β taken equal to zero, y_i reducing $\hat{y}$ but if $\beta = \frac{\bar{y}}{\bar{x}}$, so $\hat{y} = \bar{y} + \beta(\bar{X} - \bar{x}) = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \hat{\bar{Y}}$

REFERENCES

[1] Bain,L.J. and Engelhardt, M., "Introduction to probability and mathematical statistics", 2 ed., Duxbury Press, California, 1992.

[2] Cochran, W. G., "Sampling techniques", 3 ed., John Wiley and Sons, Inc., New York,1977.

[3] Efron, B. and Tibshirani, R.J., "An introduction to the bootstrap", Chapman and Hall, New York, 1993.

[4] Everitt, Brian., "An R and S-Plus Companion to multivariate analysis, Springer"., Amerika, 2004.

[5] Hardle, W.," Smoothing techniques with implementation in S", Springer Verlag, 1990.

[6] Hardle, W., Liang, H and Gao, J, "Partially linear models", Springer Verlag, Berlin, 2000.

[7] Haryatmi, Sri., "Metode Statistika Multivariat", Universitas Terbuka, Karunika, Jakarta, 1988.

[8] Jhonson, Richard. and Wichern, Dean., "Applied Multivariat Statistical Analysis", Pearson Educational International, Amerika, 2002.

[9] Rencher, Alvin., "Linier Model in Statistics", Wiley series in probability and Statistics, Canada, 2000.

[10] Rosadi, Dedi., "Analisis Ekonometrika dan runtun Waktu Terapan", Penerbit Andi, Yogyakarta, 2011.

[11] Pamungkas, Tubagus, "Estimasi dan inferensi model regresi semiparametrik proses produksi, 2012.

[12] Wibisono, Yusuf., "Metode Statistik", Gadjah Mada University Press, Yogyakarta, 2005.