Analisi Statistica per le Imprese

Prof. L. Neri

Dip. di Economia Politica e Statistica

4. La prospettiva dell'efficienza interna

Obiettivo

Ricerca dell'efficienza interna:
si tratta di determinare il corretto utilizzo delle risorse, dato un volume di produzione atteso.

Secondo il BS

Gli interventi da eseguire dovranno legare efficienza interna, aumento della produzione/ricavi, investimenti e ricerca della soddisfazione del cliente. Quindi secondo la visione rappresentata dal sistema descritto, l'efficienza interna è il primo step di una stima simultanea dell'intero sistema.

Formalmente

Consideriamo la prima delle relazioni del sistema dove la variabile X può rappresentare grandezze come:

- spese per il personale
- spese per le materie prime
- spese per gli investimenti (capitale)

ovvero ciò che riguarda i "fattori produttivi" utili per ottenere un livello di produzione prefissato.
Vediamo quindi quali metodi statistici possono essere applicati per il raggiungimento dei nostri obiettivi

Deterministic relationship

A deterministic relationship f between a dependent variable Y and an independent variables X is expressed as:

$$Y = f(X) \tag{1}$$

In the most simple case, when the relationship is linear and $k = 1$, it is commonly expressed as follows:

$$Y = \beta_0 + \beta_1 X \tag{2}$$

This relationship can be represented graphically on a Cartesian Coordinate System as a line with intercept β_0 and slope β_1.

Stochastic relationship

However deterministic relationships are very rare due to randomness. The “true” relationship between a dependent variable Y and an independent variables X is approximated by the stochastic regression model as follows:

$$Y = f(X) + \varepsilon \tag{3}$$

The term ε is supposed to incorporate the size of the approximation error. The introduction of such element identifies the stochastic relationship. If f is a linear function, the regression model is linear and is expressed as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{4}$$

The terms (β_0, β_1) are the regression parameters or model coefficients.

Regression Analysis

Regression analysis is a method to study functional relationships between variables. The relationship is expressed as an equation or model that links the dependent (fitted or predicted) variable to one or more independent (predictor) variable

Example

Si supponga di aver classificato i clienti di una banca secondo due caratteri: reddito medio disponibile e frequenza mensile di ricorso al servizio Bancomat. La relazione che si vuole verificare è se la frequenza mensile di ricorso al servizio Bancomat Y (variabile dipendente), vari al variare del reddito X . Observing a sample of size n we will have n observations for each variable.

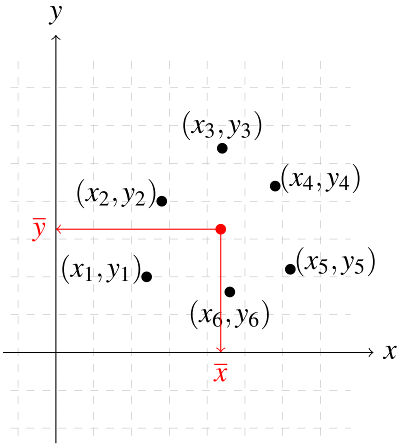
The relationship between the dependent variable and the independent variable is expressed by the following linear model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{5}$$

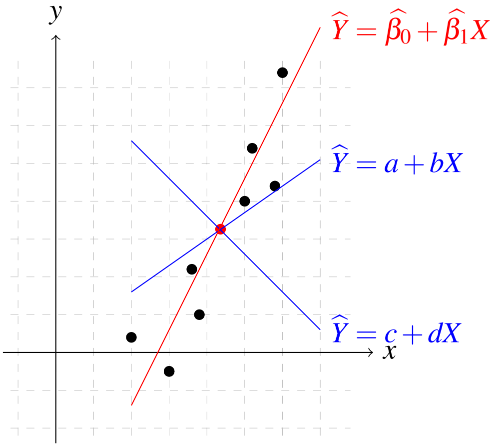
The parameters or coefficients of regression are (β_0, β_1) and ε is the random component of the model. For each observation the model can be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \; i = 1, \ldots, n \tag{6}$$

Scatter plot



Scatter plot



Optimal Parameters

One must then find the equation that best approximates the observations with coordinates (X,Y). This equation is specified by the following linear model:

$$\hat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X \tag{7}$$

Hence one must find the optimal parameters $(\widehat{\beta}_0, \widehat{\beta}_1)$.

Ordinary Least Squares (OLS)

The method of Ordinary Least Squares (OLS) minimizes the following auxiliary function:

$$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0+\widehat{\beta}_1 X\right)^2 \tag{8}$$

The function is minimized for the following parameter values:

$$\widehat{\beta}_1=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}=\frac{\sum_{i=1}^{n}x_i y_i}{\sum_{i=1}^{n}x_i^2} \tag{9}$$

$$\widehat{\beta}_0=\overline{Y}-\widehat{\beta}_1\overline{X} \tag{10}$$

Ordinary Least Squares (OLS)

$$x_i = \left(X_i - \overline{X}\right) \tag{11}$$

$$y_i = \left(Y_i - \overline{Y}\right) \tag{12}$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{13}$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{14}$$

Ordinary Least Squares (OLS)

A stochastic model is much more realistic then a deterministic model when the observations are not perfectly aligned. The introduction of the ε introduces difficulties but enhances the results that are more useful and deeper in meaning.

First consideration: Justification

How can we justify the introduction of the stochastic component?

1. There may be errors in the model
2. The number of covariates is limited
3. Randomness from empirical sampling
4. Measurement errors

Ordinary Least Squares (OLS)

Second consideration: Random Variable

The introduction of the error term allows the identification of Y as a random variable (r.v.). This implies that every value expressed as a function of Y is also a r.v.

Ordinary Least Squares (OLS)

Third consideration: Assumptions

1. The functional relationship must be linear
2. The independent variable has a deterministic nature
3. The error term has a null expected value $E[\varepsilon_i] = 0$
4. The error term is homoskedastic $Var[\varepsilon_i] = \sigma^2$
5. The error terms are not autocorrelated $Cov[\varepsilon_i \varepsilon_j]_{\forall i \neq j} = 0$

Credible assumptions

1. seems restrictive, but many non linear relationships can be expressed as linear relationships of the parameters through appropriate transformations.
2. is possibly the most unrealistic in socio-economic analysis, but guarantees important properties by $E\left[X_i\varepsilon_i\right] = X_i E\left[\varepsilon_i\right] = 0$
3. guarantees that the most likely error value is zero, as in this case the mean value is also the modal value.
4. can be unlikely in "cross section" observations, and can give serious difficulties. We will focus on this later.
5. can be unlikely in "time series" observations, and can give serious difficulties.

Gauss Markov Theorem

We will now examine the properties of the OLS estimators of the unknown parameters of the regression function. The estimates are obtained for both (Y,X) from n observations extracted with probabilistic sampling from a target population. If we extracted a new sample from the same population we would most likely obtain different results, hence the parameters estimates would be different. For this reason we can say that the estimates are tied to the r.v.
When one writes $\widehat{\beta}$ one means:

1. The slope of the regression line, estimated from a set of observations
2. the estimator that has a specific probability distribution

Properties

Given assumptions 1-5, OLS estimators are the Best Linear Unbiased Estimators (BLUE). This means that they are the most efficient (minimum variance) not systematically distorted estimators within the class of linear estimators. Here we just prove that it is linear and unbiased.

Linear estimator

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{15}$$

$$\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \sum w_i y_i \Rightarrow w_i = \frac{x_i}{\sum x_i^2} \tag{16}$$

$$\sum w_i = 0 \sum w_i x_i = 1 \tag{17}$$

Similarly it can be proven for β_0.

Unbiased estimator

$$\widehat{\beta}_1 = \sum w_i y_i = \sum w_i \left(Y_i - \overline{Y}\right) = \sum w_i Y_i - \underbrace{\overline{Y}\sum w_i}_{0} \tag{18}$$

$$\widehat{\beta}_1 = \sum w_i \left[\beta_0 + \beta_1 X_i + \varepsilon_i\right] + 0 = \underbrace{\sum w_i \beta_0}_{0} + \beta_1 \underbrace{\sum w_i X_i}_{1} + \sum w_i \varepsilon_i \tag{19}$$

$$\widehat{\beta}_1 = 0 + \beta_1 + \sum w_i \varepsilon_i \tag{20}$$

$$E\left[\widehat{\beta}_1\right] = E\left[\beta_1 + \sum w_i \varepsilon_i\right] = \beta_1 + \sum w_i \underbrace{E\left[\varepsilon_i\right]}_{0} = \beta_1 \tag{21}$$

Similarly it can be proven for β_0.

Let us add to the hypothesis 1-5 the following:
The error term has Normal distribution
so:$\varepsilon_i \rightarrow N(0, \sigma^2)$
so: $y_i \rightarrow N(\beta X_i, \sigma^2)$

OLS estimator distribution

As $\widehat{\beta}$ is a weighted average of y and the y are normally distributed, $\widehat{\beta}$ is also normally distributed.

$$\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma_\varepsilon^2}{\sum x_i^2}\right) \tag{22}$$

$$\widehat{\beta_0} \sim N\left(\beta_0, \sigma_\varepsilon^2 \frac{\sum X_i^2}{N \sum x_i^2}\right) \tag{23}$$

OLS estimators coincide with Maximum Likelihood (ML) estimators.
The Central Limit Theorem supports that even if the y where not to be normally distributed, under a very nonrestrictive conditions, the asymptotic distribution would still hold.

Residual Variance Estimator

This estimator, called the residual variance estimator, can be obtained through Maximum Likelihood computations (here omitted).

$$s^2 = \widehat{\sigma^2} = \frac{\sum \widehat{\varepsilon}_i^{\,2}}{n-2} = \frac{\sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta_1 X_i}\right)^2}{n-2} \tag{24}$$

$$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i \quad \text{(Residual)} \tag{25}$$

The residual variance is an unbiased and consistent estimator of the variance of the error term. The denominator is $n-2$ because that is the number of degrees of freedom which are calculated as the number of observations minus the required parameters in the computation. In this case to calculate this estimator one needs both (β_0, β_1).

Further observations

- $Var\left(\widehat{\beta}_1\right)$ is a direct function of $Var\left(\varepsilon_i\right)$
  - $\Rightarrow$strongly variable errors lower precision and reliability
- $Var\left(\widehat{\beta}_1\right)$ is an indirect function of the variability of X_i
  - $\Rightarrow$ covariates concentrated in a close interval worsen the quality of $\widehat{\beta}_1$

OLS Estimators standard error

Once we have estimated the variance of the stochastic term in the regression model, we may substitute it in the OLS estimator's variance to obtain the standard errors (s.e.) of estimators.

$$s^2_{\widehat{\beta}_1} = \frac{s^2}{\sum x_i^2} \tag{26}$$

$$s^2_{\widehat{\beta}_0} = s^2 \left(\frac{\sum X_i^2}{n \sum x_i^2} \right) \tag{27}$$

- Pindyck R. and Rubinfeld D., 1998 “Econometrics Models and Economic Forecasts” (4-th Ed.)
- Bracalente, Cossignani, Mulas, 2009 “Statistica Aziendale” sec.4.1