Analisi Statistica per le Imprese

Prof. L. Neri

Dip. di Economia Politica

4.2 Inferenza nel modello di regressione lineare semplice

Inference requirements

The Normality assumption of the stochastic term ε is needed for inference even if it is not a OLS requirement.
Therefore we have:

$$\varepsilon_i \sim N\left(0, \sigma^2\right) \tag{1}$$

Interpretation

The interpretation of the Confidence Interval (CI) is fixed to a level α.
If we sample more than a set of observations, each set would probably give a different OLS estimate of β and therefore different CI. $(1-\alpha)\%$ of these intervals would include β, and only $(\alpha)\%$ of the sets would deviate from β by more than a specified Δ.

Standardize a Gaussian variable

As we have previously seen:

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\varepsilon^2}{\sum x_i^2}\right) \tag{2}$$

We can standardize such statement such that:

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma_\varepsilon^2 / \sum x_i^2}} \sim N(0,1) \tag{3}$$

The t-distribution

However σ^2 in unknown, and it must be estimated by s^2.

$$\frac{N(0,1)}{\chi^2_\nu} = \frac{\frac{\left(\widehat{\beta}_1-\beta_1\right)\sqrt{\sum x_i^2}}{\sigma_\varepsilon}}{\sqrt{\frac{(\nu)s^2}{\frac{\sigma_\varepsilon}{\nu}}}} = \frac{\left(\widehat{\beta}_1-\beta_1\right)\sqrt{\sum x_i^2}}{s^2} = \frac{\widehat{\beta}_1-\beta_1}{s_{\widehat{\beta}_1}} \sim t_\nu \tag{4}$$

Where t_ν is a t distribution with $\nu = n-2$ degrees of freedom.

CI limits

Hence the CI for β at a $(1-\alpha)\%$ can be calculated as follows:

$$Pr\left(-t_{\frac{\alpha}{2}} < t_{n-2} < t_{\frac{\alpha}{2}}\right) = 1-\alpha \tag{5}$$

$$Pr\left(\underbrace{\widehat{\beta}_1 - t_{\frac{\alpha}{2}} s_{\widehat{\beta}_1}}_{\text{lower limit}} < \beta_1 < \underbrace{\widehat{\beta}_1 + t_{\frac{\alpha}{2}} s_{\widehat{\beta}_1}}_{\text{upper limit}}\right) = 1-\alpha \tag{6}$$

The CI specifies a range of values within which the true parameter is credible to lie. The level of likelihood is specified by the choice of α.

Null and Alternative Hypothesis

Given a starting conjecture (null hypothesis), considered true as working hypothesis, we may verify the Hypothesis evaluating the discrepancy between the sample observations and what one would expect from the null hypothesis. If, given a specific level of significance α, the discrepancy is significant, the null hypothesis is rejected. Otherwise it cannot be statistically rejected. It is important to notice that you may never conclusively "accept" the null hypothesis (Falsification Theory).

Hypothesis Testing Steps

- Step 1. Fix a significance level α
- Step 2. Specify the null hypothesis
- Step 3. Specify the alternative hypothesis
- Step 4. Calculate the test statistic
- Step 5. Determine the acceptance/rejection region

Hypothesis Testing: $\beta_1 = 0$

We want to test whether there is a linear relationship between X and Y at a generic $\alpha = 0.05$ level of significance, in other words we want to test the significance of the X variable in the model. The hypothesis will then be:

$$H_0 : \beta_1 = 0 \tag{7}$$

$$H_A : \beta_1 \neq 0 \tag{8}$$

Hypothesis Testing: $\beta_1 = 0$

The test statistic as before defined:

$$\frac{\widehat{\beta}_1 - 0}{s_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} \Rightarrow t_{n-2} \tag{9}$$

Hence we reject the null hypothesis iff:

$$\left| \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} \right| > \underbrace{t_{\frac{\alpha}{2};n-2}}_{\text{critical region}} \tag{10}$$

Hypothesis Testing: $\beta_1 = 0$

The Golden Rule

When n is large, the t-distribution tends to a normal distribution, hence if we choose a 5% significance level as above we can use an approximate rule of thumb and reject the hypothesis whenever the test statistics is greater the 2.

$$\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} > 2 \tag{11}$$

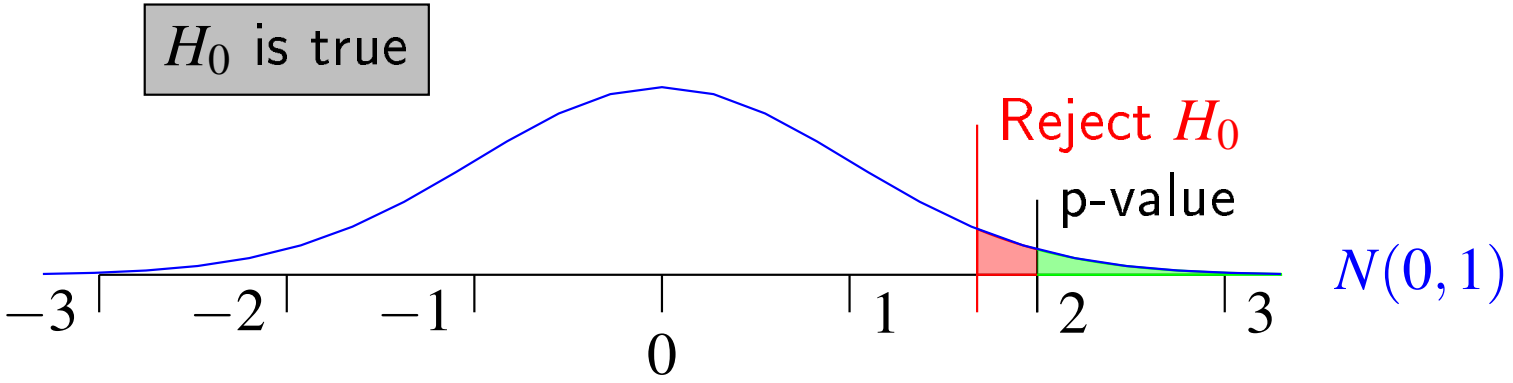
Hypothesis Testing: $\beta_1 = 0$

P-value and Significant levels

The p-value is a probability. The rule is: we reject the null hypothesis if p-value$< \alpha$

P-value example

Suppose you fix $\alpha = 0.5$, therefore $\alpha/2 = 0.025$, and the p-value associated to the coefficient $\beta_1$is 0.001. It means that effectively there is a 1% chance that the relationship emerged randomly and a 99% chance that the relationship is real. Graphically we have.



Hypothesis Testing: $\beta_1 = c$

Similarly if we are testing for a specific constant value named c:

$$H_0 : \beta_1 = c \tag{12}$$

$$H_A : \beta_1 \neq c \tag{13}$$

Hence the rejection criterion is:

$$\left| \frac{\widehat{\beta}_1 - c}{s_{\widehat{\beta}_1}} \right| > \underbrace{t_{\frac{\alpha}{2};n-2}}_{\text{critical region}} \tag{14}$$

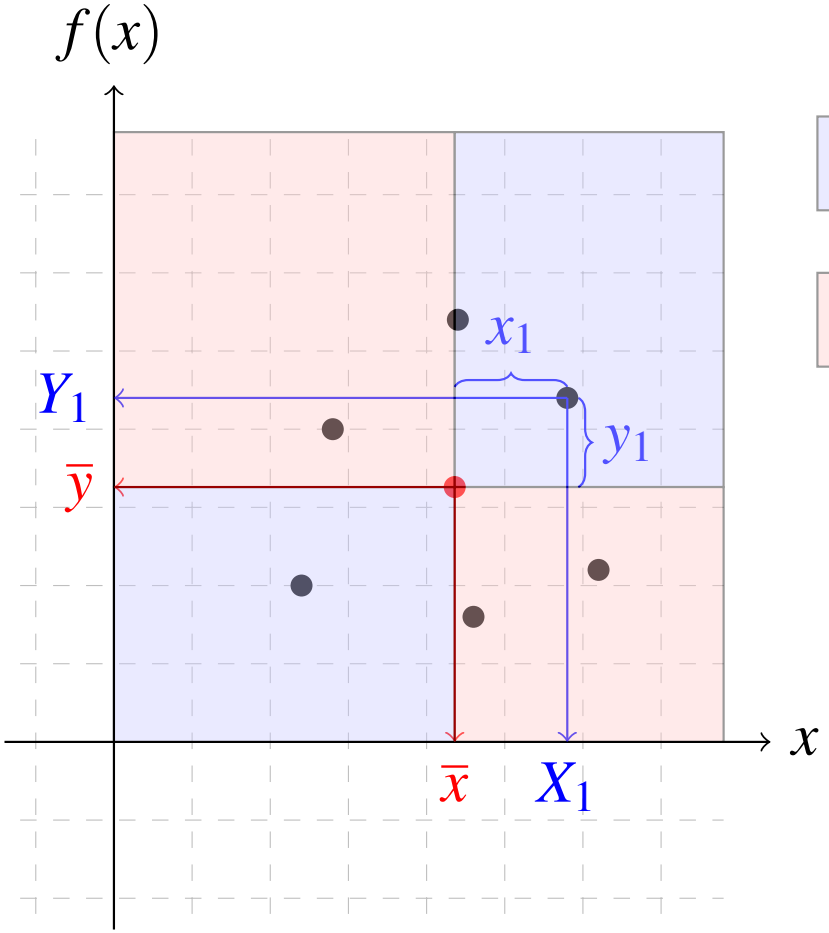
Keeping in mind the possibility of a Normal approximation for large samples.

The meaning of β_1

The parameter β_1 expresses how much change in Y follows a unitary change in X on average.

- if $\beta_1 > 0$ an increase in X is proportional to the increase in Y (direct relationship)
- if $\beta_1 < 0$ an increase in X is inversely proportional to the increase in Y (indirect relationship)

Regression and Correlation



$x_i y_i > 0 \quad 1^{st}$ and 3rd quadrant

$x_i y_i > 0 \quad 2^{nd}$ and 4th quadrant

Linear relationship

The $\sum x_i y_i$ function measures the intensity of the linear relationship between X and Y.
When this amount is expressed as mean observation contribution it is called covariance.

$$Cov(X,Y) = \frac{1}{n}\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \frac{\sum x_i y_i}{n} \tag{15}$$

The covariance is a measure that depends on the measurement scalar. An alternative that is a relative index is the Pearson correlation coefficient.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{16}$$

Relationship and Dependence

One must not however mistake the Pearson coefficient with the regression coefficient β_1 of a simple linear regression, as they have different formulas and different meanings. They can be linked by the following formula:

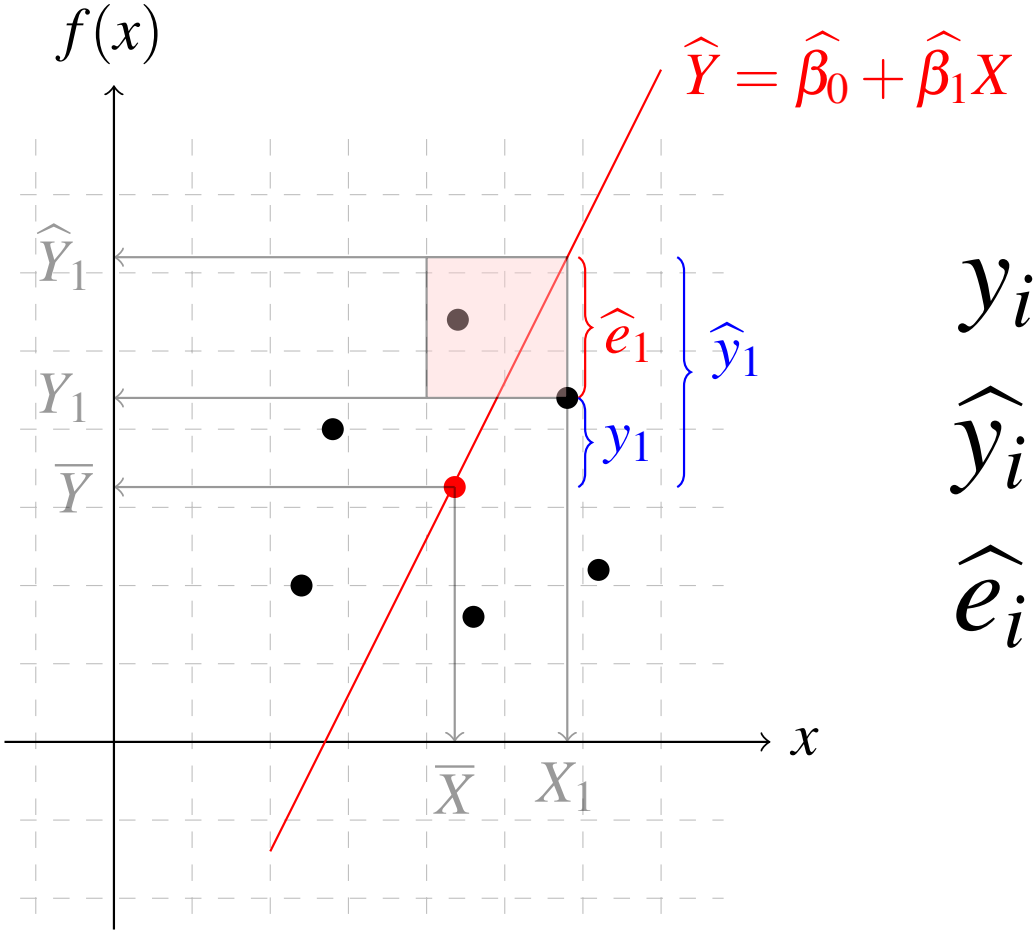
$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x}\frac{s_x}{s_y}\frac{1}{s_x} = \hat{\beta}_1 \frac{s_x}{s_y} \tag{17}$$

Clearly the meaning also differs.
The Pearson coefficient (r) measures the linear bidirectional relationship between X and Y (i.e. Y and X).

Linear relationship

- The regression coefficient β_1 measures the linear dependence of Y from X. Given a linear relationship, the dependence may be from $X \to Y$ or $Y \to X$. In regression analysis one must "ex ante" decide which dependence to investigate.
- Correlation analysis in not concerned with causal links. Regression analysis is based of causal links.

Residuals squared



$$y_i = Y_i - \overline{Y}$$
$$\widehat{y}_i = \widehat{Y}_i - \overline{Y}_i$$
$$\widehat{e}_i = Y_i - \widehat{Y}_i$$

Residual properties

$$\sum \widehat{\varepsilon}_i = 0; \quad \sum \widehat{\varepsilon}_i X_i = 0 \tag{18}$$

$$\sum \widehat{\varepsilon}_i = \underbrace{\sum y_i}_{0} - \widehat{\beta}_1 \underbrace{\sum x_i}_{0} = 0 \tag{19}$$

$$\sum \widehat{\varepsilon}_i X_i = \sum x_i \left(y_i - \widehat{\beta}_1 x_i \right) = \sum x_i y_i - \widehat{\beta}_1 \sum x_i^2 \tag{20}$$

$$\sum \widehat{\varepsilon}_i X_i = \sum x_i y_i - \frac{\sum x_i y_i}{\sum x_i^2} \sum x_i^2 = 0 \tag{21}$$

Sum of squares

$$Y_i - \overline{Y} = \left(Y_i - \widehat{Y}_i\right) + \left(\widehat{Y}_i - \overline{Y}\right) \tag{22}$$

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum\left(Y_i - \widehat{Y}_i\right)^2 + \sum\left(\widehat{Y}_i - \overline{Y}\right)^2 + \underbrace{2\sum\left(Y_i - \widehat{Y}_i\right)\left(\widehat{Y}_i - \overline{Y}\right)}_{0} \tag{23}$$

$$\underbrace{\sum \left(Y_i - \overline{Y}\right)^2}_{\text{TSS}} = \underbrace{\sum \left(Y_i - \widehat{Y}_i\right)^2}_{\text{RSS}} + \underbrace{\sum \left(\widehat{Y}_i - \overline{Y}\right)^2}_{\text{ESS}} \tag{24}$$

1. TSS = Total Sum of Squares
2. RSS = Residual Sum of Squares
3. ESS = Explained Sum of Squares

R^2: coefficient of determination

$$\text{TSS} = \text{RSS} + \text{ESS} \Rightarrow \frac{\text{RSS}}{\text{TSS}} + \frac{\text{ESS}}{\text{TSS}} = 1 \tag{25}$$

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{26}$$

The coefficient of determination R^2 expresses the proportion of total deviance explained by the regression model of Y on X. As one cannot explain more then all the existing deviance we can write as follows:

$$\max(ESS) = TSS \Rightarrow 0 \leq R^2 \leq 1 \tag{27}$$

R^2 properties

- $R^2 = 0$ implies that the explicative contribution of the model is null, hence the deviance is completely expressed by the random component.
- $R^2 = 1$ implies that all the observations lie on the regression line, hence the whole variability is explained by the model.
- All intermediate cases imply that the higher/lower the value of R^2 the more/less of variability is expressed by the choice of the model. A model with an $R^2 = 0.80$ implies that 80% of the variability is explained by the chosen model.
- The R^2 is said to be a "fitting index" as it measures the adaptiveness of the chosen model to the specified dataset.

R^2 and r

$$R^2 = \frac{ESS}{TSS} = \left(\widehat{\beta}_1 \frac{s_x}{s_y}\right)^2 = (r)^2 \tag{28}$$

Hence the coefficient of determination is equal to the coefficient of correlation squared. Another equally simple and efficient relationship is can be obtained by:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \widehat{\varepsilon}_i^2}{\sum y_i^2} \tag{29}$$

ANalysis Of VAriance

The total variability decomposition shows the error component and the model component.

$$\text{TSS} = \text{RSS} + \text{ESS} \Leftrightarrow \sum y_i^2 = \sum \widehat{\varepsilon}_i^2 + \sum \widehat{y}_i^2 \tag{30}$$

We also know that:

$$\text{ESS} = \sum \widehat{y}_i^2 = \widehat{\beta}_1^2 \sum x_i^2 \tag{31}$$

$$\frac{\left(\widehat{\beta}_1 - \beta_1\right)\sqrt{\sum x_i^2}}{\sigma_\varepsilon} \sim N(0,1) \tag{32}$$

ANalysis Of VAriance

As the square of a Normal distribution is a Chi-squared distribution:

$$\frac{\left(\widehat{\beta}_1 - \beta_1\right)^2 \sum x_i^2}{\sigma_\varepsilon^2} \sim \chi_1^2 \tag{33}$$

Omitting here the proof for:

$$\frac{\sum e_i^2}{\sigma_\varepsilon^2} \sim \chi_{n-2}^2 \tag{34}$$

The ratio of χ^2 divided by the d.f. is known to be an F distribution, as follows:

$$\frac{\left(\widehat{\beta}_1 - \beta_1\right)^2 \sum x_i^2}{\sum e_i^2 / (n-2)} \sim F_{1;n-2} \tag{35}$$

ANalysis Of VAriance

One may then verify the null hypothesis $H_0 : \beta_1 = 0$ with respect to $H_A : \beta_1 \neq 0$ with the following test statistic:

$$\frac{\left(\widehat{\beta}_1\right)^2 \sum x_i^2}{\sum e_i^2 / (n-2)} = \frac{\mathrm{ESS}/1}{\mathrm{RSS}/(n-2)} \sim F_{1,n-2} \tag{36}$$

A strong linear relationship between X and Y will determine a high valued test statistic, which supports the model, and reject the null hypothesis. Therefore:

Anova table

	Deviance	d.f.	Correct variance estimate
Model	ESS$=\sum \widehat{y}_i^2$	1	$\widehat{\beta_1^2} \sum x_i^2 / 1$
Residual	RSS$=\sum \widehat{\varepsilon}_i^2$	$(n-2)$	$\sum \widehat{\varepsilon}_i^2 / (n-2)$
Total	TSS$=\sum \widehat{y}_i^2 + \sum \widehat{\varepsilon}_i^2$	$(n-1)$	

Table 1: Anova table

Forecast

Often the estimated regression model is used to predict the expected Y value (i.e $\widehat{Y}$) which corresponds to a specific value of the X variable in this case indicated with X_κ.

$$\widehat{Y_\kappa} = \widehat{\beta}_0 + \widehat{\beta}_1 X_\kappa \tag{37}$$

The standard error of such value is equal to:

$$s.e.\left(\widehat{Y_\kappa}\right) = s\sqrt{1 + \frac{1}{n} + \frac{\left(X_\kappa - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}} \tag{38}$$

Forecast

The $1-\alpha$ confidence intervals for such value are then specified as:

$$\widehat{Y_\kappa} \pm t_{(\alpha/2;n-2)} s.e.\left(\widehat{Y_\kappa}\right) \tag{39}$$

We may notice that the value of the s.e. increases proportionally to the distance between X_κ and $\overline{X}$, therefore taking distant forecasts will worsen the quality of the estimate. It can also happen that the linear relation breaks down for very high or low X_κ values, as it is limited to the observe scatter plot. In this case taking a forecast outside the range of the observed values can be misleading.

Numerical example

Applicazione del modello di regressione lineare semplice tratto da De Luca (1996)

Consideriamo l'andamento del valore degli impieghi bancari nelle regioni italiane in funzione del numero di imprese dell'industria e dei servizi.

Y = valore impieghi bancari (miliardi di vecchie lire)

X = numero di imprese dell'industria e dei servizi (migliaia di unità).

Obiettivo: predire l'entità di questi ultimi in rapporto all'evoluzione presumibile del numero di imprese considerate

esempio numerico

la retta stimata con R fornisce il seguente risultato:

$$y = -12861,67 + 247,99x \tag{40}$$

questo significa che se il numero di imprese cresce di 1000 unità gli impieghi bancari aumentano di circa 248 miliardi di lire.

- Pindyck R. and Rubinfeld D., 1998 "Econometrics Models and Economic Forecasts" (4-th Ed.)
- Bracalente, Cossignani, Mulas, 2009 "Statistica Aziendale" sec.4.1
- De Luca A.,1996 "Marketing Bancario e Metodi Statistici Applicati", vol. 1, Franco Angeli.