Regression models and panel data
Part II: Linear regression for cross-sections

Rodolfo Metulini

✉ rmetulini@unisa.it

Department of Economics and Statistics (DISES) - University of Salerno

Outline

Normal stochastic variables

- We work with stochastic variables, or "variabili aleatorie" (va)
- When considering a sample of observations (e.g. the vector of the annual income of n workers, the vector of the annual gross domestic product for n countries, etc..), statistically, we consider each value of the vector a realization from a stochastic variable, so...
- y_i (income of the $i-$th worker) is a realization from the va Y
- Often we assume that Y is a *va* with a normal distribution, so $Y \sim N(\mu, \sigma)$, where $E(Y) = \mu$ and $V(Y) = \sigma^2$
- As a matter of fact, the normality assumption in the (linear regression) model is, however, on the residuals
- Probability density function: $f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}$
- Important: each $i-th$ observation is a realization of a stochastic variable. With notation $y \sim iid\ N(0, \sigma)$ we mean that all the elements in the vector y are independent and presents the same distribution

Independence between two (or more) variables

- In linear regression models we work with two or more stochastic variables, let say, Y and X
- Considering two **events** A and B, if $P(A \cap B) = P(A)P(B)$, A and B are not dependent ($A \perp\!\!\!\perp B$). Also $P(A \mid B) = P(A)$.
- When talking about **variables** (let say Y and X) with a distribution, we say that, if $E(Y \mid X) = E(Y)$, it follows that $X \perp\!\!\!\perp Y$.
- **Linear independence** between two variables can be measured on the vector of realizations x and y. If there exists scalars a_1, a_2 such that $a_1 x + a_2 y = 0$, where 0 is a vector of zeros, X and Y are dependent. The same can be generalized for more than two variables (e.g. $X_1, X_2, .., X_p$).
- A more elegant way to check for linear independence is to measure the rank of the design matrix of dimension nxp $X = [X_1, X_2, ..., X_p]$. If the rank is p, the p variables are each others independent.

Covariance and correlations

- A measure used for linear dependence among two variables is the covariance (and the correlation)

$$Cor(X, Y) = 0 (uncorrelation) \nrightarrow X \perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \rightarrow Cor(X, Y) = 0$$

- $Cov(X, Y) = E(XY) - E(X)E(Y)$ (for X and Y two stochastic variables)
- $Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ (for sample realizations from X and Y)
- $Cor(X, Y) = \frac{Cov(X,Y)}{\sqrt{var(X)var(Y)}}$, where $var(X) = \frac{\sum_{i=1}^{n} (x_1 - \bar{x})^2}{n}$
- Warning: Cor and β of the regression model are not comparable

The regression function (i)

- The most general formulation for the causal relation between one or a set of independent variables and a dependent variable can be expressed, deterministically, as:

$$Y = f(X_1, X_2, ..., X_p)$$

where f may be a linear or non linear function.

- If we just consider one independent variable:

$$Y = f(X)$$

- If we consider a linear relation, the most simple reads as:

$$Y = \beta_o + \beta_1 * X$$

The regression function (ii)

- Switching to a "non deterministic" or "stochastic" formulation, the causal relation between Y and X may be expressed as:

$$Y = f(X) + \varepsilon$$

- ε (also called "disturbance") is a stochastic variable with $E(\varepsilon) = 0$ which permits to take into account in the model the effects of all not considered variables that may have an effect on Y.
- E.G. let consider a group of bank customers with the same income (where income is the variable X). It is unlikely that all of them will have exactly the same savings $Y = f(X)$.

The regression function (iii)

- The regression function is deterministic on X and "stochastic" on ϵ:

$$E(Y \mid X) = E(f(X) + \varepsilon) = E(f(X)) + E(\varepsilon) = f(X)$$

- Given a sample of realizations (empirical observations) $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, it is possible to explicit the regression function for each of the $i-$realization of the sample (generally, of dimension n):

$$y_i = f(x_i) + \varepsilon_i, \forall i = 1, 2, ..., n$$

Specification (i)

- The linear specification reads as:

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i, \forall i = 1, 2, ..., n$$

with

$$E(Y \mid X) = \beta_0 + \beta_1 * x$$

- The expected value for the dependent variable y_i ($\hat{y}_i$) is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$$

where

$$\epsilon_i = y_i - \hat{y}_i$$

is the sample realization of the stochastic variable ε on the $i-$th sample realization.

Model assumptions (i)

- 1) **Linearity on the parameters**: The relation between Y and X is linear

  Warning: the following formulations are linear in the parameters:

  $y_i = \beta_0 + \beta_1 * log(x_i) + \varepsilon_i$

  $y_i = \beta_0 + \beta_1 * \frac{1}{x_i} + \varepsilon_i$.

  This is not linear in the parameters:

  $y_i = \beta_0 + \beta_1^2 * x_i + \varepsilon_i$

- 2) **Zero mean of the disturbances**: $E(\varepsilon_i) = 0, \forall i = 1, ..., n$
- 3) **Homoschedasticity in the disturbances**: $Var(\varepsilon_i) = \sigma^2 < -\infty, \forall i = 1, ..., n$

Model assumptions (ii)

- 4) **Independence between the disturbances**: $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$
- 5) **Normality of the disturbances**: $\varepsilon_i \sim N(0, \sigma^2)$

  The error terms are normally distributed
- 6) **X deterministic**: The indipendent variable X is known without errors.

Ordinary Least Squares (i)

- The most used estimation method for determining an estimate for parameters β_0 and β_1 is that of Ordinary Least Squares (OLS).
- The objective is that of minimize a function, called Q:
- $\min_{\beta_0,\beta_1} Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 * x_i)^2$
- The solution of the problem will be:

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

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ordinary Least Squares (ii)

In summary:

- We denote with β_0 e β_1 the true (unknown) parameters expressing the causal relation between X and Y.
- We instead denote with $\hat{\beta}_0$ and $\hat{\beta}_1$ the corresponding OLS estimates based on the sample realizations of the variables X and Y.

Properties of OLS estimates (i)

- **Linearity**: $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the sample realizations $y_1, y_2, ..., y_n$ and $x_1, x_2, ..., x_n$
- **Unbiasness**:

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

- **Efficiency**: Among all estimators for β made with a linear combination of all the sample realizations, OLS estimates are those with smaller variance.

Properties of OLS estimates (ii)

- **Consistency**: For sample size n that tends to ∞, the estimates converge to the value that the estimator is designed to estimate (with 0 variance).
- **Best Linear Unbiased Estimators (BLUE)**: OLS estimates for β_0 and β_1 and those with smaller variance, among all the possible unbiased estimators.
- **Asymptotic normality**: OLS estimates for β_0 and β_1 distribute normally as n tends to ∞.

An unbiased estimator for the variance

- SS_ε stays for "Sum of squares of errors" and it is a measure of deviance of the disturbances.
- $\frac{SS_\varepsilon}{n-2}$ is the variance
- $Var(\varepsilon) = \frac{SS_\varepsilon}{n-2}$ is an unbiasness estimator for σ^2:

$$E(Var(\varepsilon)) = E(\frac{SS_\varepsilon}{n-2}) = \sigma^2$$

where:

$$SS_\varepsilon = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Variance decomposition

- The linear regression model admits the following variance decomposition:

$$Var_{total} = Var_{model} + Var_{residual}$$

or, deviance decomposition:

$$Dev_{total} = Dev_{model} + Dev_{residual}$$

which can be equivalently expressed as:

$$SS_T = SS_R + SS_\varepsilon$$

or:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The index of linear determination, R^2

- $R^2 = \frac{SS_R}{SS_T} = \frac{Var_{model}}{Var_{total}}$
  or, equivalently:

$$R^2 = 1 - \frac{SS_\varepsilon}{SS_T} = 1 - \frac{Var_{residual}}{Var_{total}}$$

- R^2 admits value in the range [0,1]
- $R^2 = 0$ when the model is completely **inadequate** to explain the relation among X and Y, according to the sample data;
  $R^2 = 1$ when the model is completely **adequate** to explain the relation among X and Y, according to the sample data.

The multiple linear regression model

- Let assume the following relation is in place: $y_i = x_i\beta + \varepsilon_i$
- y_i is the $i-$th sample realization from the stochastic variable Y.
- x_i is a 1 x p vector containing the sample realizations of the p independent variables for the $i-$ realization: $x_i = (x_{i,1}, ..., x_{i,p})$.
- β is the vector of regression coefficients of dimension p x 1.
- ε_i is the stochastic variable for the disturbances.

The multiple linear regression model in matrix form (i)

- Let suppose to have n units with $n > p$.
- Let $\mathbf{Y} = (y_1, ..., y_n)'$ be the $nx1$ vector of the dependent variable for the n observations.
- Let $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ be the $nx1$ vector of the disturbances for the n observations.
- Let $\mathbf{X}$ be the n x p matrix with the realizations of the n independent variables.

The multiple linear regression model in matrix form (ii)

- In a matricial formulation, the multiple linear regression model reads as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

- If the model contains the intercept, the first column of $\mathbf{X}$ is a vector of ones.

Assumptions of the multiple linear regression model

1. $\mathbf{X}$ is a deterministic matrix with rank p.
2. $\mathrm{E}(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.
3. $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.
4. $\varepsilon_i \sim N(0, \sigma^2)$.

We can summarise hypotesis 2 and 3 with:

$$E(\varepsilon\varepsilon^T) = \sigma^2 I_n$$

where I_n is an identity matrix of dimension n x n

Rank of a matrix (i)

- Let A be a rectangular matrix of dimension m x n.
- From A we can extract a number of squared submatrices, called "minors", by removing a row or a column.
- Let "order" be the number of columns (or rows) of that submatrix.
- For each minor we can compute the determinant.
- The rank of A is the largest order of the minors such that the determinant is different from 0.

Discussion assumption 1

- If rank of $\mathbf{X}$ is p, it follows that all the independent variables are linearly independent.
- In other words, it does not exist an independent variable that can be written as a linear combination of the others.

Discussion assumption 2

- From assumption 2 it follows that:

$$E(y_i \mid x_i) = f(x_i), \forall i$$

and:

$$Var(y_i \mid x_i) = \sigma^2, \forall i$$

Estimates of β and σ (i)

- Let suppose assumptions 1–3 are in place.
- Let $\mathbf{Y}^T = (y_1, ..., y_n)$ be the 1 x n vector of the sample realizations from the v.c. $\mathbf{Y}$.
- We can obtain an estimation of the vector β with OLS:

$$S = \sum_{i=1}^{n} (y_i - x_i\beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

- After calculations:

$$S = \mathbf{Y}^T\mathbf{Y} - \mathbf{X}^T\beta^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^TX\beta =$$
$$\mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

Estimates of β and σ (ii)

- To obtain $\hat{\beta}$ we apply the partial derivates of S in terms of each β_i. In matrix notation:

$$\frac{\delta S}{\delta \beta} = -2\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\beta + \mathbf{X}^T\mathbf{X}\beta = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta$$

- By setting the previous equation to 0 and by solving for β we have that:

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\beta$$

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

Estimates of β and σ (iii)

- Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ the vector of estimated y
- Let $\epsilon = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ the vector of empirical disturbances.
- The variance of the residuals is:

$$\hat{\sigma}^2 = \frac{\epsilon^T \epsilon}{n - p}$$

Unbiasness of the estimators

- We reminds that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$
- Considering the equation for the multiple linear regression $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ we have that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T(\mathbf{X}\beta + \varepsilon))$$

- $E(\hat{\beta})$ is:

$$E(\hat{\beta}) = E[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T(\mathbf{X}\beta + \varepsilon)] =$$

$$\mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\beta] + E[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T)\varepsilon)] =$$
$$\beta + (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T)E(\varepsilon) = \beta$$

- $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

Coefficients interpretation (i)

- Let suppose to have just two independent variables: X_1 and X_2.
- The relation among $\mathbf{Y}$ and the two independent variables is:

$$E(\mathbf{Y} \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- β_1 represents the effect on $\mathbf{Y}$ of a unitary variation on X_1, by taking X_2 constant.
- Viceversa for β_2
- β_0 represents the expected value of $\mathbf{Y}$ when both X_1 and X_2 are 0.

Coefficients interpretation (ii)

- Let suppose X_1 varies of an amount equal to δX_1 and that X_2 remains constant.
- In light of a variation of X_1, $\mathbf{Y}$ varies as well:

$$\mathbf{Y} + \delta\mathbf{Y} = \beta_0 + \beta_1(X_1 + \delta X_1) + \beta_2 X_2$$

- We obtain that :

$$\delta\mathbf{Y} = \beta_1 \delta X_1$$

from which it follows:

$$\beta_1 = \frac{\delta\mathbf{Y}}{\delta X_1}$$

Adjusted R^2

- By including one more variable to the regression model, the R^2 increases, even when the independent variable is not significant.
- Fo such a reason is preferable to use the adjusted R^2 in the multiple linear regression model:

$$R^2_{adj} = 1 - \frac{n-1}{n-p}\frac{SS_\varepsilon}{SS_T}$$

- The correction factor $\frac{n-1}{n-p}$ is always larger than 1, so $R^2_{adj} < R^2$.

The t-student test

- We want to test the hypothesis that the $j-th$ β coefficient (β_j) is not significantly different from 0, under the assumption σ^2 is unknown.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- $$t = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\hat{\sigma}^2 v_{jj}}}$$

  where v_{jj} is the $j-th$ element of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$

- t, under the null hypothesis, is distributed as a t-student with $n-p$ degrees of freedom.
- If empirical t is larger than tables' t or smaller than minus tables' t, we reject the null hypothesis.

The F statistic (i)

- Let suppose we want to test the following system of hypothesis:

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_{p-1} = 0 \\ H_1 : \beta_j \neq 0 \text{ for at least one } j,\ j = 1, ..., p-1 \end{cases}$$

- In other words, under the null hypothesis, all the independent variables (except for the intercept) do not have any significant effect on $\mathbf{Y}$.

The F statistic (ii)

- Let $SSE_{restricted}$ the residual sum of squares from the restricted model (i.e. that with just the intercept).
- $SSE_{unrestricted}$ the residual sum of squares from the unrestricted model (i.e. that with all the independent variables).
- The F statistics is:

$$F = \frac{(SSE_{restricted} - SSE_{unrestricted})/(p-1)}{(SSE_{unrestricted})/(n-p)}$$

- The F statistic (F_{obs}) must be compared with the theoretical F (F_{th}) on the tables. If $F_{obs} > F_{th}$, we reject the null hypothesis.

References

- Greene, W. H. (2000). Econometric analysis 4th edition. International edition, New Jersey: Prentice Hall, 201-215.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.