GENERAL THEORETICAL NOTIONS ON UNIVARIATE REGRESSION

Prof. univ. dr. Constantin ANGHELACHE
Academia de Studii Economice din Bucureşti/Universitatea „Artifex„ din Bucureşti
Prof. univ. dr. Ion PARTACHI
Academia de Studii Economice a Moldovei, Chisinau
Conf. univ. dr. Mădălina-Gabriela ANGHEL
Universitatea „Artifex„ din Bucureşti
Drd. Gyorgy BODO
Drd. Radu STOIAN
Academia de Studii Economice din Bucureşti

Abstract:
In this article, the authors started from the fact that in general, the concept of conditional probability and the conditional linear probability in terms of orthogonal projections are common to the crowd of linear functions. Against this background, a presentation on the main conditionality involved in univariate regression was conducted. Thus, linearity, uncolinerity and conditional normality are presented and demonstrated. At the same time, homoscedasticitate conditioning is highlighed. Further, the presentation of conditions linearity and homoscedasticitate is based on the concept of error highlighting to be considered for univariate regression. Further, it points out that the estimation is performed by the method of least squares parameter which reduces to beta estimation. Another element considered and clarified concerns the replacement of the probability through the distribution probability sampling, that is subject to minimization criterion.

Finally, it highlights the fact that ordinary least squares estimator $\hat{\beta}$, of β has a minimum variation in the family of all linear unbiased estimators's of β '.

Key words: Conditional probability, integrable function, exogenous variable, univariate regression, orthogonal projection

Introduction

In general, are common the concepts of conditional probability and linear conditional probability – or linear regression – in terms of orthogonal projections the purposes of rule L^2, on the set of integrable functions of any vector *z*, indicated by L^2(z), and on the set of linear functions of *z*, indicated by $L^2(z)$. We will aplly the concept of regression and linear regression to specific models to study estimation and testing problems.

We will mention some general concepts of linear regression. Lets consider a statistic model $M_n = \left\{X^n, \Theta, P_n^\theta\right\}$ where $X^n \subset \Re^{nm}$ is the sample domain n, Θ is the parametre domain, and P_n^θ is the family distributions of the sample. Let be $x \in X^n$ a finite series $(x_i)_{i=1,...,n}$ with $x_i = (y_i, z_i)'$ where $y_i \in \Re, z_i \in \Re^q, q = m-1$. We will consider only the case in which the size p of y_i, is equal to one. The observations $x_1,...,x_n$ are independent in the purpose of distribution P_n^θ for all θ and distributed by the same distribution of the probability $Q^\theta : P_n^\theta = [Q^\theta]^{\otimes n}$. Moreover, we assume that for all $1,...,n, y_i$ şi z_{ij} for $j=1,..., q$ are square integrable random variables, eg. belong to $L^2(z)$.

The regression model presentation in terms of probability it is, in the linear case, the same as the one presented by Spanos (1986), who called it linear regression model in contrast

to the Gaussian linear model, defined by the normal equation $y = X\beta + u$. For the last one, the exogenous variables seem to be decisive. Indeed, the linear regression model is based on general probabilistic arguments and the Gaussian linear model is just a particular case. In this presentation, rather a general one of the regression model, we can deepen the study based on same author works.

Literature review

Regarding the linear regression model, a more rigorous proof of the OLS equivalence and the method of moments can be found in Gouriéroux and Mon-Fort (1996a). Restricted regression model has been studied by numerous authors, especially Gouriéroux and Monfort (1996a, Volume 2), Greene (1990), Spanos (1986), Judge, Griffiths, Hill, Lutkepohl, and Lee (1985), and Judge, Hill, Griffiths, Lutkepohl, and Lee (1988). To show that β_n is BLUE, Judge, Hill, Griffiths, Lutkepohl and Lee (1988) can be consulted in particular. For an estimation, you can consult Spanos (1986), Greene (1990), and Judge Hill, Griffiths, Lutkepohl and Lee (1988). The main issues for log linear model can be found in Greene (1990). To demonstrate the convergence of σ_n^2 to σ^2 Monfort (1982) can be consulted. In a comparison analysis of test procedures of Wald, Rao and LR in case that the null hypothesis can be expressed in the form $R\beta = r$,it may be deepen the study of authors Judge, Griffiths, Hill. Lutkepohl, and Lee (1985) and Spanos (1986).

Regarding nonlinear parametric regression we are pointing for further study Bierens (1994), Gouriéroux and Monfort (1996a, Volume I), Greene (1990, Chapter11), Spanos (1986); and for the linear restriction test - Bierens (1994).

The part relating to the models incorrectly specified it can be deepen by studing the works of White (1980), Florens, Ivaldi şi Larribeau (1996), Gallant and White (1988) and White (1994).

Methodology and data

We are interested in the conditional probability of the form

$$E^{\theta}(y_i|z_i) = g(z_i, \beta).$$

For all *i*, where β is a function of θ that we suppose is finite. We consider this vector parameter estimation problem β, assuming that g is known. Definig the random vector *u* in the domain $\Re^p$ and $u = (u_1,...,u_n)'$ for all *i*, we obtain:

$$u_i = y_i - E^{\theta}(y_i|z_i). \qquad (1)$$

The properties that will lead to the next model:

$$y_i = g(z_i, \beta) + u_i \quad i = 1, \ldots, n. \qquad (2)$$

The choise of g, or more generally the choice of conditioning or design determines the type of model to be considered. In this way, if the domain is restricted to the set $L^{*2}(z)$. In the field of linear functions of *z*, we obtain:

$$g(z_i, \beta) = EL^{b}(y_i|z_i) = \beta' z_i.$$

Linear regression

First, we consider the assumptions that allow us to specify what we call linear regression model. Let us remember that $y_i \in \Re$ and $z_i = (z_i,..., z_{iq})'$ for all *i=1,…, n* with the property that $z_i = 1$, to which we will return later.

First we will assume the linearity condition

- For all i=1, . . . , n, we have:

$$E^{\theta}(y_i|z_i) = EL^{\theta}(y_i|z_i) = \beta' z_i = \sum_{j=1}^{q} \beta_j z_{ij}$$

, cu $\beta = (\beta_1,...,\beta_q)'$, in addition $E^{\theta}(z_i z_i')$ is reversible.

We can deduce from this assumption the general form of linear regression equation, using the expression for any term u_i resulting from the relation (1), as follows:

$$y_i = \beta' z_i + u_i \qquad i = 1, \ldots, n.$$

We define vectors *n x 1 - y* and *u* by $y = (y_1,..., y_n)'$ and $u = (u_1,...,u_n)'$ and matrix *Z* *n* x *q* by function:

$$Z = (z_{ij})_{i,j} = (z_1, \ldots, z_n)'.$$

The matrix of the regression equation is $y = Z\beta + u$.

By construction, the term u_i possess a number of properties. For all *i=1,…,n,* we get:

$$E^{\theta}(u_i|z_i) = 0$$

We note that the independence *xi,* implies that $E^{\theta}(u_i \mid z_i) = E^{\theta}(u_i \mid Z)$ and so

$$E^{\theta}(u|Z) = 0.$$

Then we deduce:

$$E^{\theta}(u_i z_i|z_i) = z_i E^{\theta}(u_i|z_i) = 0. \qquad (3)$$

This second property represents the conditional orthogonality between u_i *şi* z_i. Also, these properties remain valid in terms of the marginal probabilities:

and $E^{\theta}(u_i) = E^{\theta}\left[E^{\theta}(u_i|z_i)\right] = 0$

$$E^{\theta}(u_i z_i) = 0$$

For *i=1,…,n,* that is the fundamental equation to be estimated.

- The next conditioning indentified is the assumption of non-coliniarity:

$Rank(Z) = q \quad (n > q).$

This can be written in the following equivalent ways $Rank(Z'Z) = q, \det(Z'Z) \neq 0$ or *Z'Z* reversible. This assumption is equivalent in terms of distributional assumption that the sample matrix variation $E^{\theta}(z_i z_i')$ is reversible which does not depend on *i.*

Another conditioning is that of homoscedasticity, respectively:

$$Var^{\theta}(y_i|z_i) = \sigma^2, \text{ for all } i = 1, \ldots, n.$$

We can infer immediately that $Var^{\theta}(y_i \mid z_i) = \sigma^2$, for all *i=1,…, n.* Furthermore, in order of all *i* şi *j* cu *i≠ j,* we obtain:

$$
\begin{aligned}
Cov^{\theta}(u_i, u_j|Z) &= E^{\theta}(u_i u_j|Z) \\
&= E^{\theta}\left[\left(y_i - E^{\theta}(y_i|z_i)\right)\left(y_j - E^{\theta}(y_j|z_j)\right)|Z\right] \\
&= E^{\theta}\left[\left(y_i - E^{\theta}(y_i|Z)\right)\left(y_j - E^{\theta}(y_j|Z)\right)|Z\right] \\
&= E^{\theta}(y_i y_j|Z) - E^{\theta}\left[y_i E^{\theta}(y_j|Z)|Z\right] - E^{\theta}\left[y_j E^{\theta}(y_i|Z)|Z\right] \\
&\quad + E^{\theta}\left[E^{\theta}(y_i|Z)E^{\theta}(y_j|Z)|Z\right] \\
&= 0.
\end{aligned}
$$

As long as y_i and y_j are conditional independent Z, we deduced:

$$E^{\theta}(u_i u_j|Z) = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

This property of the error term can be written in the realtionship form:

$$E^{\theta}(u_i u_j) = E^{\theta}\left[E^{\theta}(u_i u_j|Z)\right] = \begin{cases} \sigma^2 & Daca\ i = j \\ 0 & Daca\ i \neq j \end{cases}$$

Or

$$Var^{\theta}(u) = E^{\theta}(uu') = Var^{\theta}(u|Z) = E^{\theta}(uu'|Z) = \sigma^2 I_n$$

The last conditioning is id the normality conditioning, when is normally distributed i = 1,..., n.

Conditionings l, 3, and 4 can be resumed by:

$$y_i|z_i \sim i.i.N(\beta' z_i, \sigma^2).$$

For all $i = 1$,... ,n, implying that $u_i|z_i \sim i.i.N(0, \sigma^2)$ more precisely

$$u_i \sim i.i.N(0, \sigma^2).$$

This last property (conditioning) summarizes the basic principles of linear regression models in a large number of econometrics manuals that specify the model in this way, starting with the error term and derived from it the assumptions of orthogonality, linearity and homoscedasticity. In this case, the conditioning is specified.

The estimation by ordinary least squares is reduced to the estimation of the vector parameter β. We will use here the results of the previous chapter about the notion of the best aproximation based on the purposes of the rule L^2. The estimator you get you get with this method is the ordinary last square *(OLS)* of β.

The β estimator it is obtain as a solution to the problem of minimizing the next function over λ, ie:

$$E^{\theta}\left[\left(y_i - \sum_{j=1}^{q} \lambda_j z_{ij}\right)^2\right] = E^{\theta}\left[\left(y_i - \lambda' z_i\right)^2\right],$$

From here we get the following simple equations, respectively

$$E^{\theta}\left[z_i(y_i - \lambda' z_i)\right] = 0. \qquad (5)$$

The (4) equation defines β as a solution to a minimization problem, as:

$$\phi(x_i, \lambda) = (y_i - \lambda' z_i)^2.$$

The first order conditions (5) form a simple equation of time, fixing

$$\psi(x_i, \lambda) = z_i(y_i - \lambda' z_i).$$

Replacing the probability in sampling the probability distribution calculated using the empirical distribution, transforms to minimize:

$$D(\lambda_1, \ldots, \lambda_q) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{q} \lambda_j z_{ij} \right)^2 = \sum_{i=1}^{n} (y_i - \lambda' z_i)^2 \qquad (6)$$

to $\lambda_1,...,\lambda_q$, or in terms of minimizing the matrix:

$$D(\lambda) = (y - Z\lambda)'(y - Z\lambda).$$

to $\lambda = (\lambda_1,...,\lambda_q)$.

Primele condiţii ale ordinii problemei minimizării sunt:

$$\frac{\partial D(\lambda)}{\partial \lambda} = -2Z'y + 2Z'Z\lambda = 0. \qquad (7)$$

This can be rewritten as:

$$Z'(y - Z\lambda) = 0$$

or

$$\sum_{i=1}^{n} z_i(y_i - \lambda' z_i) = 0.$$

This last relationship allows us to get the expression for the moment estimator $\hat{\beta}_n$, referred to herein as the smallest fours ordinary estimator β, ie:

$$\hat{\beta}_n = (Z'Z)^{-1}Z'y = \left[\sum_{i=1}^{n} z_i z_i' \right]^{1} \sum_{i=1}^{n} z_i y_i. \qquad (8)$$

Following the logic notation, this estimator should be noted with $\hat{\lambda}_n$. We will use the notation $\hat{\beta}_n$ to remain in the demonstration system assumed002E

The second condition is satisfied for orders:

$$\frac{\partial^2 D(\lambda)}{\partial \lambda \partial \lambda'} = Z'Z$$

which is a positive semi-defined matrix, such $\hat{\beta}_n$, it is a minimum.

Considerând, pentru toate *i = 1,...,n,*

$$\hat{y}_i = \hat{\beta}_n' z_i$$

and

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

We can define the vectors $\hat{y}$ and $\hat{u}$, both with the size *n* x 1, by:

$$\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)' \quad \text{şi} \quad \hat{u} = (\hat{u}_1, \ldots, \hat{u}_n)'$$

Under conditionings 1-3, when σ^2 in unknown, it can be estimated by:

$$\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_n' z_i \right)^2$$

şi

(9)

$$\sigma^2 = E^{\theta}(u_i^2) = E^{\theta}\left[(y_i - \beta' z_i)^2\right].$$

Totuşi, vom prefera un estimator diferit, care rămâne imparţial, adică:

$$\widehat{\sigma}_n^2 = \frac{1}{n-q}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_n' z_i\right)^2, \qquad (10)$$

$\widetilde{\sigma}_n^2$ it can be written as:

$$\widehat{\sigma}_n^2 = \frac{1}{n-q}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{1}{n-q}(y-\widehat{y})'(y-\widehat{y}) = \frac{1}{n-q}\widehat{u}'\widehat{u}.$$

where $\hat{\sigma}_n^{\ 2}$ is equal to:

$$\begin{aligned}\widehat{\sigma}_n^2 &= \frac{1}{n-q}(y - Z\widehat{\beta}_n)'(y - Z\widehat{\beta}_n)\\ &= \frac{1}{n-q}(y'y - y'Z(Z'Z)^{-1}Z'y)\\ &= \frac{1}{n-q}y'M_Z y\end{aligned}$$

and

$$M_Z = 1 - Z(Z'Z)^{-1}Z'$$

We will consider the conditional density of y_i has the form:

$$f(y_i|z_i, \beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(y_i - \beta' z_i)^2}{2\sigma^2}\right\}.$$

The maximization problem in this case is:

$$\max_{\lambda,\rho} E^{\theta}\left[\ln f(y_i|z_i, \lambda, \rho)\right].$$

MLE of β and σ^2, we express through $\widetilde{\beta}_n$, şi $\widetilde{\sigma}_n^2$ satisfies the relationship:

$$(\widetilde{\beta}_n', \widetilde{\sigma}_n^2) = \arg\max_{\lambda,\rho}\frac{1}{n}\sum_{i=1}^{n}\ln f(y_i|z_i, \lambda, \rho) =$$

$$= \arg\max \ln l_n(y|Z, \lambda, \rho),$$

where l_n is the given expresion of the probability

$$l_n(y|Z, \lambda, \rho) = \prod_{i=1}^{n} f(y_i|z_i, \lambda, \rho) \quad = \frac{1}{\rho^n(2\pi)^{n/2}}\exp\left\{-\sum_{i=1}^{n}\frac{(y_i - \lambda' z_i)^2}{2\rho^2}\right\}$$

The estimators are then derived, namely

$$\frac{\partial}{\partial\lambda}\ln l_n(y|Z, \lambda, \rho) = 0$$

and

$$\frac{\partial}{\partial\rho}\ln l_n(y|Z, \lambda, \rho) = 0$$

This is equivalent to

$$\frac{\partial}{\partial\lambda}\sum_{i=1}^{n}\left(y_i - \lambda' z_i\right)^2 = \frac{\partial}{\partial\lambda}(y - Z\lambda)'(y - Z\lambda) = 0 \qquad (11)$$

and

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \lambda' z_i)^2 \qquad (12)$$

In terms of the relationship (11). From the relation (12) we can derive the expression for $\tilde{\sigma}_n^2$ and we get:

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{\beta}_n' z_i)^2 = \frac{n}{n-q}\hat{\sigma}_n^2. \qquad (13)$$

The properties of the small samples will be highlighted starting with the finite sample $\hat{\beta}_n$. $\hat{\beta}_n$ it can be written in the form *Ay*, cu $A = (Z'Z)^{-1}Z'$, and thus linear in *y*. More,

$$\hat{\beta}_n = (Z'Z)^{-1}Z'(Z\beta + u) = \beta + (Z'Z)^{-1}Z'u,$$

So that,

$E^{\theta}(\hat{\beta}_n | Z) = \beta$; therefore $\hat{\beta}_n$ is an impartioal estimator of β. Conditional variance is given by:

$$Var^{\vartheta}(\hat{\beta}_n | Z) = E^{\theta}((Z'Z)^{-1}Z'u | Z) = \sigma^2(Z'Z)^{-1}.$$

Thus, we get to Gauss-Markov theorem, which states:

„Ordinary least squares estimator $\hat{\beta}_n$ of β has a minimum variation in the family of all linear unbiased estimators of β".

To demonstrate this, we will consider a different linear estimator of β, denoted $\tilde{\beta}_n$, of form $\tilde{\beta}_n = Cy$, where *C* is a matrix *q x n* , of form:

$$C = D + (Z'Z)^{-1}Z'$$

In the above equation, *D* has the dimension *q x n.* Suppose this new estimator is unbiased, that is:

$$E^{\vartheta}(\tilde{\beta}_n | Z) = E^{\vartheta}\left[\left(D + (Z'Z)^{-1}Z'\right)(Z\beta + u) | Z\right] = DZ\beta + \beta$$

Means that means $DZ = 0$ and as a consequence:

$$\tilde{\beta}_n = \beta + (D + (Z'Z)^{-1}Z')u.$$

Computing the variation of $\tilde{\beta}_n$, using the $DZ = 0$ constraint, and also $Var^{\theta}(u \mid Z) = \sigma^2 I_n$, we obtain:

$$\begin{aligned} Var^{\theta}(\tilde{\beta}_n | Z) &= Var^{\vartheta}((D + (Z'Z)^{-1}Z')u | Z) \\ &= \sigma^2 DD' + \sigma^2(Z'Z)^{-1} \\ &= \sigma^2 DD' + Var^{\vartheta}(\hat{\beta}_n | Z) \\ &\geq Var^{\vartheta}(\hat{\beta}_n | Z) \end{aligned}$$

Due to the fact that *D'D* is a positive, semi-defined matrix, results that in case of a finite sample $\hat{\beta}_n$, is the *best linear unbiased estimator* (BLUE) of β.

Taking into consideration the properties of $\hat{\sigma}_n^2$, for proving that is an estimator, we start from the equation:

$\hat{u} = y - \hat{y} = M_Z y = M_Z(Z\beta + u) = M_Z u$., where $M_Z Z = 0$, we obtain the result:

$$\widehat{\sigma}_n^2 = \frac{1}{n-q} u' M_Z u.$$

The conditional probabilities resulting from this calculation is:

$$E^\theta(\widehat{\sigma}_n^2|Z) = \frac{1}{n-q} E^\theta(u' M_Z u|Z) = \frac{1}{n-q} E^\theta(tr(u' M_Z u)|Z) = \frac{1}{n-q} tr\left[M_Z E^\theta(uu'|Z)\right] = \frac{\sigma^2}{n-q} tr M_Z,$$

From where and we can conclude that:

$$E^\theta\left(\widehat{\sigma}_n^2|Z\right) = \sigma^2,$$

If this is a scalar, then $tr(a) = a$; if.*A* and *B* are two arrays of similar size, and consequently *tr(AB) = tr(BA);*

It also results that:

$$tr(M_Z) = tr(I_n) - tr(Z(Z'Z)^{-1}Z') =$$
$$= n - tr((Z'Z)^{-1}Z'Z) = n - q$$

And we can conclude that:

Se poate arăta de asemenea că

$$Var^\theta\left(\widehat{\sigma}_n^2|Z\right) = \frac{2\sigma^4}{n-q}.$$

We can conclude that the maximum probability estimator is unbiased.

Conclusion :

This article highlights the main general theoretical concepts on univariate regression. In this context, presents conditionalities that must be taken into account in the construction and use of univariate regression in practical economic analysis. In this analysis we've annalysed the presentation of the regression model in terms of probability, in accordance with the theories expressed by a number of econometrics specialists (eg Spanos) is a linear regression model defined by the mathematical equation Y = Xβ + ε. In this model, the considered exogenous variable is crucial. Estimation of β parameter in such a function is performed by ordinary least squares method. The obtained estimator from this method is, in fact, the estimator of least ordinary squares (OLS) of β. In the presentation made, we believe that the maximum likelihood estimator is unbiased. Although univariate regression analysis is rarely used in social-economic, it is recomanded to use when, in fact, is a form of nonlinear regression.

References

1. Andrei, T., Bourbonais, R. (2008). *Econometrie,* Editura Economică, Bucureşti
2. Ang, A. and G. Bekaert (2007). *Stock return predictability: Is it there?*, Review of Financial studies 20 (3), pp. 651–707
3. Anghelache, C. (2016). *Econometrie teoretică – Ediţia a II-a revizuită,* Editura Artifex, Bucureşti
4. Anghelache, C., Manole Alexandru (2016). *Utilizarea modelului de regresie în analiza corelaţiei dintre situaţia monetară şi balanţa de plăţi / The use of regression model in analysing the correlation between the monetary situation and the balance of payments*, Romanian Statistical Review, Supplement, no.7, pg. 24-29 / 30-42
5. Anghelache, C., Sacală, C. (2016). *Multiple linear regression used to analyse the corelation between GDP and some variables*, Romanian Statistical Review, Supplement, no.9, pp. 94-99
6. Anghelache, C., Anghel, M.G. (2014). *Using the regression model in the analysis of financial instruments portfolios*, Procedia Economics and Finance, pp. 324-329, Volume 10/2014
7. Anghelache, C. (2011). *Elemente de econometrie aplicată*, Editura Artifex, Bucureşti
8. Bosq, D. (2012). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, Springer Science & Business Media
9. Ghysels, E. (2001). *The Econometric Analysis of Seasonal Time Series,* Cambridge University Press

10. Lohr, S.L. (2007). *Comment: Struggles with Survey Weighting and Regression Modeling*, Statistical Science, Vol. 22, No. 2, pp.175–178
11. Newey, W., Powell, J. (2003). *Instrumental variable estimation of nonparametric models*, Econometrica, pp. 1565–1578
12. Pecican, E.Ș. (2009). *Econometria pentru... economiști. Econometrie - teorie și aplicații, ediția a treia*, Editura Economică, București
13. Pesavento, E., Rossi, B. (2006). *Small–sample Confidence Interevals for Multivariate Impulse Response Functions at Long Horizons*, Journal of Applied Econometrics 21(8), pp. 1135–1155