Gross Product Simulation with pooling of Linear and Nonlinear Regression Models

Ahmad Flaih[1], Abbas Abdalmuhsen[1], Ebtisam Abdulah[1], Srini Ramaswamy[1]

[1] University of Arkansas at Little Rock
Little Rock, AR 72204, USA
{anflaih, akabdalmuh, ekabdulah, sxramaswamy}@ualr.edu

Abstract. This paper discusses the problem of decision support systems in the organization. The procedure (linear combination) developed with the aim to combine some predicted results obtained with simulation of linear and nonlinear regression models (experts), multiple regression model, nonparametric regression model, and semi parametric regression model. This adjustment procedure enforce some statistical characteristics like the expected value of the gross production rate based on Cobb-Douglas production function is unbiased for the actual value, and the total weights(importance) of all models(experts) is equal to one. We used modeling and simulation techniques to generate our data and to apply the procedure.

Keywords: Regression Models, Linear combination, Cobb-Douglas production function, Predict of the gross product.

1 Introduction

In economics, the relationship between the production (outputs) and the production elements (inputs) known as production function. Usually decision makers in the organization and companies used some statistical and economics methods to support the production policy. Cobb-Douglas (CD) production function is one of the most useful tools to support the production policy, [1] and [9] used the CD function and found that public investment has a large contribution to production, [7] argued that public-sector capital has no effect on production after controlling for location characteristic, [2] and [5] pursued the insignificance of public capital. All of the literature above assumes a Cobb-Douglas (CD) production function to estimate productivity of inputs. One property of the CD functional form is that elasticity of an input is the estimated slope coefficient when the output and inputs are measured in logs, [6] used nonparametric perspective, which allows elasticities to vary across location and time, [15] use the CD functions proposed by [9]. [2] and [6] considers the gross product(GP) as responding variable, public capital (PUC), private capital (PRC), the employment rate (EMP), and the unemployment rate (UEPR) as explanatory variables, to predict of the gross product by using linear and nonlinear regression models. [13] they have propsed linear pooling method to combine the probability forecasts and proved the propsed method gave them more accurate results.

In this paper we used the same variables and methodology to predict gross production, but we use linear combination method to combine the results of the three experts (multiple linear regression, non-linear regression, and semi-parametric) as

J. Barjis, M.M. Narasipuram, G. Rabadi, J. Ralyté, and P. Plebani (Eds.):
CAiSE 2010 Workshop EOMAS'10, Hammamet, Tunisia, pp. 69-76, 2010.

input variables, we choose these three types of models because we think there is linear relationship, non-linear relationship and mixture relationship between gross production and the predictor variables. It is based on our hypothesis that the linear combination method will give us more accurate predictive gross production since we believe this is a better way to take into account the results of multiple models accompanied by an appropriate weighting scheme. Furthermore, we prove that this method gives an unbiased estimator to the gross production. Our contribution is that we use linear combination method to pool three regression models for predicting the gross production.

Linear combination is linear pooling of some sets of ordered pairs (importance, gross production), we have explored the prediction associated with different regression model can be combined into one final model via weighted linear combination will give more accurate results. Mathematically linear combination of the sequence $y_1, y_2, ..., y_n$ each with mean μ is:

$Y = \sum_{i=1}^{n} a_i y_i$, where a_i is the weight of y_i ,and

$$\mathrm{E}(Y) = \sum_{i=1}^{n} a_i \mathrm{E}(y_i) = a_1\mu + a_2\mu + ... + a_n\mu \quad \textbf{(1)}$$

The $\mathrm{E}(Y) = \mu$ if $\sum_{i=1}^{n} a_i = 1$

2 The Models

Regression analysis is a technique used in data analysis; we use regression technique to predict the value of the response (dependent) variable given any value of the predictor (independent) variable. A general regression model is [15]:

$$y_i = E(y_i | x_i) + e_i \quad \textbf{(2)}$$

Where i=1, 2... n denoting an observation of a subject. y_i is the response variable and x_i is a $k \times 1$ vector of independent variables. E ($y_i | x_i$) is the expectation of y_i conditional on x_i, and e_i is the error term. In this paper we will use the following types of regression model:

2.1 Parametric Regression Model:

In this model, it is assumed that y_i is linearly related with x_i, so we can say it is linear regression model:

$$E(y_i | x_i) = \alpha + x_i' \beta$$

Thus a linear regression model is written as:

$$y_i = \alpha + x_i' \beta + e_i \quad \textbf{(3)}$$

Where α is the intercept and β is $k \times 1$ vector of parameters. Under Gauss-Markov assumptions, the estimators of α and β are Best Linear Unbiased Estimators (BLUE), and can be estimate by using Ordinary Least Squares method (OLS).

2.2 Nonparametric Regression Model

If we do not know the data generating process, it is very unlikely that a linear regression (parametric) model is exactly the appropriate model specification, so the estimators α and β are not BLUE (best linear unbiased estimator). Instead of making assumptions for the functional form of E ($y_i \mid x_i$), nonparametric regression methods do not require any presumptions about the underlying data generating process [10]. Let

$$y_i = m(x_i) + e_i \tag{4}$$

Where m (.) is some function of unspecified functional form. Some basic assumptions about m (.) are commonly made. Many methods have been devised to estimate the regression function m (.), but we will just consider a simple but effective estimate known as the Kernel regression estimate. Suppose we have a random sample (x_1, y_1), (x_2, y_2) …, (x_n,y_n). The Kernel regression estimate of m(x) =E ($y \mid X=x$) is given by:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} k(\frac{x_i - x_j}{d}) y_i}{\sum_{i=1}^{n} k(\frac{x_i - x_j}{d})} \tag{5}$$

Here, K is a nonnegative symmetric function that is not increasing as its argument gets away from zero, and d is a parameter called the smoothing parameter that is selected by the user to control the amount of smoothing. The estimator in equation (4) is called the Local- Constant Least- Squares, which can be interpreted as a weighted average of y_i, where:

$$\frac{k(\frac{x_i - x_j}{d})}{\sum_{i=1}^{n} k(\frac{x_i - x_j}{d})}$$ is the weight attached to y_i

It should be noted that the product Kernel function $\mathrm{K}\ (\frac{x_i - x_j}{d})$ is the product of the Kernel function of all components of x. That is:

$$\mathrm{K}\ (\frac{x_i - x_j}{d}) = \prod_{s=1}^{k} k(\frac{x_{is} - x_{js}}{d_S}) \tag{6}$$

Where x_{js} is the s^{th} component of x_j and d_s the s^{th} component of d. The Kernel function k (.) can take several forms. In this paper we used the Gaussian Kernel function is defined as:

$$k(\frac{x_{is} - x_{js}}{d_S}) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x_{is} - x_{js}}{d_S})^2] \tag{7}$$

The smoothing parameter d is generally the most important factor when performing nonparametric regression. The bandwidth is chosen to obtain a desirable

trade-off between the bias and the variance of estimation, so we need a method that could balance the bias and variance of the resulting estimate.

We have used Leave-One-Out Cross-Validation to select the optimal bandwidth. This method is depending on the principle of selecting bandwidth that minimizes the sum squared error of the resulting estimates. We will try to find the minimum MSE, mathematically; we are trying to minimize the following function [15]:

$$\text{CV (d)} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_j(x_j)]^2 \qquad \textbf{(8)}$$

Where $\hat{m}_j(x_j)$ is the Leave-One-Out estimator of m (.) evaluated at x_j. SAS software uses this method to find the optimal bandwidth values.

2.3 Semi parametric regression model:

Nonparametric techniques are very flexible because they do not need any assumptions about functional form. However, there are cases in which the relationship between the dependent variable and some independent variables is known to be linear and the relation between the dependent variable and other independent variables remain undetermined. In this situation semi parametric models are developed to solve this problem. Semi parametric models have both parametric and nonparametric components [15].

2.3.1 Semi parametric Partially Linear Model

Consider the semi parametric partially linear model:

$$y_i = x_i'\beta + m(z_i) + e_i \qquad \textbf{(9)}$$

Where β is a $k\times 1$ vector of parameters, z_i is a $q\times 1$ vector of independent variables and e_i is an additive error, is assumed to be uncorrelated with xi and z_i. β is the parametric part of the model and the unknown function m (.) is the nonparametric part. [8] [11] proposed an estimate of β and m (z_i) as follows. Using ordinary least squares, the estimator of β is:

$$\hat{\beta} = (\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i')^{-1}\sum_{i=1}^{n}\tilde{x}_i'\tilde{y}_i \qquad \textbf{(10)}$$

Where:

$\tilde{x}_i = x_i - E(x_i \mid z_i)$ and $\tilde{y}_i = y_i - E(y_i \mid z_i)$

Once we obtain $\hat{\beta}$, the nonparametric part m (z_i) is easy to estimate. From equation (8), we get: $m(z_i) = y_i - \tilde{x}_i\beta - e_i$

Then we can get the estimator of m (z_i) as follow:

$$\hat{m}(z_i) = \frac{\sum_{i=1}^{n}(y_i - \tilde{x}_i\beta)k(\frac{z_i - z_j}{d})}{\sum_{i=1}^{n}k(\frac{z_i - z_j}{d})} \qquad \textbf{(11)}$$

Where d can be estimate similar to the nonparametric model? SAS software uses Cross-Validation to find the optimal bandwidth.

3. Simulation Study

We will generate a random sample of size 30 observations for each explanatory variables: public capital (PUC), private capital (PRC), the employment rate (EMP), and the unemployment rate (UERP). The Cobb-Douglas (CD) production function is,

$$y_i = f(PUC_i, PRC_i, EMP_i, UERP_i) \quad \textbf{(12)}$$

Where y_i =gross product, PUC_i =public capital,

PRC_i =private capital, EMP_i =employment rate (labor)

$UERP_i$ =unemployment rate.

We can rewrite the CD function as,

$$\mathrm{E}(y_i) = \beta_0 PUC_i^{B_1} PRC_i^{B_2} EMP_i^{B_3} UERP_I^{B_4} \varepsilon_i$$

The log-linear CD production function is

$$y_i = \beta_0 + \beta_1 PUC_i + \beta_2 PRC_i + \beta_3 EMP_i + \beta_4 UERP_i + \varepsilon_i \quad \textbf{(13)}$$

Further,

1- If $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$, the product function has constant returns to scale.

2- If $\beta_1 + \beta_2 + \beta_3 + \beta_4 < 1$, returns to scale are decreasing.

3- If $\beta_1 + \beta_2 + \beta_3 + \beta_4 > 1$, returns to scale are increasing.

3.1.1 Parametric model Results

We apply multi-linear regression model, covariance model, and variance component model to our generated data to find the estimated values by using SAS procedure proc reg [12] and the output are listed in Tables 1, follow:

Table 1: Estimates of output Elasticity: parametric approaches

Parametric	β_0	β_1	β_2	β_3	β_4
OLS S.E	15.089 (5.4239)	-0.0157 (0.3465)	0.29952 (0.218)	-1.709 (1.107)	-0.6721 (0.2114)
covariance model	1.87 (4.23)	0.342 (0.034)	0.005 (0.1933)	-.0127 (0.039)	-.004 (0.001)
variance component model	0.018 (0.023)	0.266 (0.020)	0.755 (0.023)	0.347 (0.052)	-0.005 (0.001)

The results of Table 1 are based on the model in [12].

3.1.2 Non Parametric Model

We used this model to find the estimated values of the intercept and elasticity values by using generated data and SAS procedure (proc gam). The output from proc gam is listed in following table, Table 2:

Table 2: Estimates of output Elasticity: Nonparametric approach

Nonparametric	β_0	β_1	β_2	β_3	β_4
Mean	14.834	-0.211	0.2409	-1.242	-0.653
S.E	(6.342)	(0.37)	(0.391)	(1.401)	(0.27)

3.1.3 Semi- Parametric Model

We used this model to find the estimates values of Elasticity by using SAS procedure (proc gam) and we considered the independent variable (Public Capital) is linearly related to the gross production. The output from proc gam is listed as follows in Table 3.

Table 3: Estimates of output Elasticity: semiparametric approach

Semiparametric	β_0	β_1	β_2	β_3	β_4
Mean	19.892	-0.459	0.288	-2.010	-0.642
S.E	(7.369)	(0.470)	(0.296)	(1.50)	(0.28)

3.2 Linear Combination Method

We will use the estimated models (experts), multiple parametric regression, nonparametric regression, and Semiparametric regression to predict o f the log-gross production, for sample of size equal to 10 observations, and by using the values of the columns(Parametric y_1,Nonparametric y_2,Semiparametric y_3) in the following table, we can estimates the weight corresponding to each expert(model) with SPSS software by using neural networks procedure and estimate the log-gross production Y as follow:

$$Y = 0.194y_1 + 0.595y_2 + 0.255y_3$$

Correlation is a measure of the association between two variables; it is a very important part of statistics. One of the most fundamental concepts in many applications is the concept of correlation. If two variables are correlated, this means that you can use information about one variable to predict the values of the other variable. The correlation matrixes between the predicted values in the table 4 as follows:

Table 4: Correlation between Predicted Values

i	y_1	y_2	y_3	Actual	Y
1	12.445	11.234	12.876	12.98	12.41
2	12.458	11.011	12.455	12.23	12.08
3	12.765	10.203	12.897	11.70	11.84
4	11.239	11.054	12.987	12.12	12.07
5	12.456	12.899	12.765	13.25	13.31
6	12.345	12.876	12.849	13.56	13.33
7	12.455	11.112	12.098	12.32	12.11
8	13.345	13.453	12.069	13.55	13.66
9	13.678	12.456	12.932	13.40	13.36
10	12.546	12.453	12.948	13.23	13.14

From Table 5 we can find that the correlation coefficient between the predicted values of the linear combination(Y) and the actual variable of gross production is (0.954) that means the linear combination method is more associated with the actual values than the expert1, expert2, and expert3.this strongest association supports the suggestion that the predicted value of the gross production which associated with the different experts that combined is better than taking the predictive values of each expert individually. The average difference is (0.261, 0.959, 0.146, 0.103) of the three regression models (parametric, non-parametric, and semi-parametric) respectively and linear combination procedure. The proposed method (linear combination) is more accurate than regression models because the difference (actual and Y) in table 4 is more less than the other models, so this means that the proposed method give us unbiased estimator.

Table 5: Correlation Matrix

	y_1	y_2	y_3	Actual	Y
y_1	1	0.407	-0.268	0.457	0.525
y_2	0.407	1	-0.149	0.935	0.981
y_3	-0.268	-0.149	1	-0.015	-0.047
Actual	0.457	0.935	-0.015	1	0.954
Y	0.525	0.981	-0.047	0.954	1

4. Conclusion

In this paper, table 5(correlation matrix) shown that the correlation coefficients between the linear combination method, parametric model, nonparametric model and semi-parametric and the actual values of the gross production are (0.954, 0.457, 0.935, and -0.015), because we believe the public capital (PUC) is linearly related to gross production, so we choose it as parametric variable and that is why the

correlation between y_3 and actual value is -0.015. Linear combination method is more associated with the actual value than the others experts (regression model), i.e. linear combination method reflects the experts views to find the most fitted predicted value to the actual value. Finally as forecasters often wish to provide an accurate predicted value, so we will use the linear combination method to predict of the gross production.

References

1. Aschauer David A. : public investment and productivity growth in the Group of Seven. 89-13, Federal Reserve Bank of Chicago (1989).
2. Baltagi,B.and Pinnoi,N. : Public capital stock and state productivity growth:Further evidence from and error component model.Empirical Economics 20,351-359 (1995).
3. Clemen, R. T. and Winkler, R. L. :"Combining Probability Distributions from Experts in Risk Analysis", Risk Analysis 19,187-203 (1999).
4. Dong Xiang "Fitting Generalized Additive Models with the GAM Procedure", SAS Institute Inc., Cary, NC 27513 (1990).
5. Garcia-Mila,Teresa,Therese.J.McGuire,and Robert H. ,Porter: The effect of public capital in state-level production function reconsidered .Review Economics and Statistics forthcoming (1996).
6. Henderson,D.J. and Kumbhakar,S. :Public and private capital productivity puzzle:A nonparametric approach,southern Economic Journal 73(1),219-232 (2006).
7. Holtz-Eakin,D. :Public-sector capital and the productivity puzzle,Review of Economics and statistics 76,12-21 (1994).
8. Li, Q. and Racine, J.S. : Nonparametric econometrics: theory and practice, Princeton University Press (2007).
9. Munnell,A. :How does public infrastructure affect regional economic performance? New England Economic Review September,11-32 (1990).
10. Par Osterholm :"Incorporating Judgment in Far Chart", Finance and Economics Discussion Series. Screen Reader Version (2006).
11. Robinson, P.M. : Root-n-consistent semiparametric regression, Econometrica 56, 931-954 (1988).
12. Ronald P. Cody and Jeffrey K. Smith : "Applied Statistics and the SAS Programming Language" 5th.ed (2006).
13. Roopesh R., and Tilmann G. : Combining Probability Factors", Technical Report no.543 (2008).
14. Schick, A. : On asymptotically efficient estimation in semiparametric models, Annals of Statistics 14, 1139-1151 (1986).
15. Xianghang S.:"Application of Nonparametric and Semi-parametric methods in Economics and Finance", PhD dissertation, University of Southern California (2009).