The XIII International Conference
"Applied Stochastic Models and Data Analysis"
(ASMDA-2009)
June 30-July 3, 2009, Vilnius, LITHUANIA

ISBN 978-9955-28-463-5
L. Sakalauskas, C. Skiadas and
E. K. Zavadskas (Eds.): ASMDA-2009
Selected papers. Vilnius, 2009, pp.90–94

ON NONLINEAR REGRESSION MODEL FOR CORRESPONDENCE MATRIX OF TRANSPORT NETWORK

Alexander Andronov[1], **Diana Santalova**[2]

Riga Technical University, Faculty of Transport and Machine Engineering,
Kalku str. 1, LV-1658 Riga, Latvia
E-mail: [1]*Aleksandrs.Andronovs@rtu.lv;* [2]*Diana.Santalova@rtu.lv*

Abstract: A nonlinear regression model for a forecasting of passenger flow between various geographical points (towns) is described. Unknown parameters are estimated using aggregated data when only the information on the departed from each town passenger's number is available. As the estimation efficiency criterion the weighted sum of residual squares is used.

Keywords: nonlinear regression, estimation, gradient method.

1. Introduction

We have n corresponding points (towns) with numbers i = 1, 2, …, n. For the point i, one is known inhabitants (citizens) number h_i and m numerical characteristics (regressors) $c_{i,j}$, j = 1, 2, …, m, those are known constants. For all pairs of the points (i,l) the distance $d_{i,l}$ between them is known as well. In addition, we know the number of the departed passengers Y_i from the point i during considered time interval, that is a random variable.

Our aim is to estimate correspondence size $Y_{i,l}$ for all pairs of points (i,l), precisely the number of the departed passengers from the point i to the point l. The matrix of $Y_{i,l}$ is to be said *the correspondence matrix*. Let us denote an estimate of $Y_{i,l}$ by $Y^*_{i,l}$. It requests that all $Y^*_{i,l}$ to be positive $\left(Y^*_{i,l} > 0\right)$ for $i \neq l$, $Y^*_{i,i} = 0$ and $Y^*_{i,l} = Y^*_{l,i}$.

As a model for the concrete correspondence (i,l) for $i \neq l$ we use

$$Y_{i,l} = \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\left(a + \left(c_{(i)} + c_{(l)}\right)\alpha + g_{(i,l)}\beta + V_{i,l}\right), \tag{1}$$

where a, $\alpha = (\alpha_1 \quad \alpha_2 \quad \ldots \quad \alpha_m)^T$ and $\beta = (\beta_1 \quad \beta_2 \quad \ldots \quad \beta_m)^T$ are unknown regression parameters, τ and θ are unknown form parameters, $c_{(i)} = (c_{i,1} \quad \ldots \quad c_{i,m})$ and $g_{(i,l)} = (c_{i,1}c_{l,1} \quad \ldots \quad c_{i,m}c_{l,m})$ are m–vector-rows, $\{V_{i,l}\}$ are independent identically distributed random variables with zero mean and unknown variance σ^2.

In addition we set $Y_{i,i} = 0$. Note that the case θ = 1 and τ = 2 corresponds to the so-called gravity model.

As a corollary of this model we get the following presentation for the number of the departed passengers from the point i:

$$Y_i = \sum_{\substack{l=1 \\ i \neq l}}^{n} Y_{i,l} = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\left(a + \left(c_{(i)} + c_{(l)}\right)\alpha + g_{(i,l)}\beta + V_{i,l}\right). \tag{2}$$

Now we must estimate unknown parameters on the basis of fixed values $\{Y_i\}$. Such a problem was considered earlier in the literature. Besides, usually the entropy approach is used for that. But there are

many obtained estimates of $Y_{i,l}^{*}$ equal to zero that is inaccessible. We use regression theory (Srivastava, 2002). For that we need to investigate the distribution and the expectation of $Y_{i,l}$.

2. Distribution analysis

We suppose that $V_{i,l}$ has normal distribution. Then $Z_{i,l} = \exp(V_{i,l})$ has the log-normal distribution (Sleeper, 2007) with characteristics

$$E(Z_{i,l}) = E(\exp(V_{i,l})) = \exp\left(\frac{1}{2}\sigma^2\right),$$
$$D(Z_{i,l}) = D(\exp(V_{i,l})) = \exp(\sigma^2)(\exp(\sigma^2)-1).$$

Therefore, for $i \neq l$

$$E(Y_{i,l}) = \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right)\exp(\frac{1}{2}\sigma^2), \tag{3}$$

$$D(Y_{i,l}) = \frac{(h_i h_l)^{2\theta}}{(d_{i,l})^{2\tau}} \exp\left(2\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right)\right)\exp(\sigma^2)(\exp(\sigma^2)-1) = \tag{4}$$
$$= (\exp(\sigma^2)-1)(E(Y_{i,l}))^2.$$

Analogous formulae have places for $\{Y_i\}$:

$$E(Y_i) = \sum_{l=1}^{n} E(Y_{i,l}) = \exp\left(\frac{1}{2}\sigma^2\right)\sum_{\substack{l=1 \\ i\neq l}}^{n} \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right), \tag{5}$$

$$D(Y_i) = (\exp(\sigma^2)-1)\sum_{l\neq i}(E(Y_{i,l}))^2. \tag{6}$$

3. The least squares estimates

We wish to use the expectation (5) for the estimation of the unknown parameters θ, τ, a, α, β and σ^2. But one cannot identify both parameters a and σ^2 simultaneously. So, let us introduce the united parameter $\tilde{a} = a + \frac{1}{2}\sigma^2$ and rewrite (5) as

$$E(Y_i) = \sum_{\substack{l=1 \\ l\neq i}}^{n} \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(\tilde{a} + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right). \tag{7}$$

As a criterion of estimates efficiency we use weighted least squares sum:

$$R(\gamma, w) = \sum_{i=1}^{n} w_i (Y_i - E(Y_i))^2, \tag{8}$$

where $\gamma = \left(\theta \quad \tau \quad \tilde{a} \quad \alpha^T \quad \beta^T\right)^T$ and $w = (w_1 \quad w_2 \quad \dots \quad w_n)^T$ is a vector of weights.
For a minimization of (7) we use the gradient method. Let

$$\nabla R(\gamma, w) = \left(\frac{\partial}{\partial\theta}R \quad \frac{\partial}{\partial\tau}R \quad \frac{\partial}{\partial\tilde{a}}R \quad \frac{\partial}{\partial\alpha}R \quad \frac{\partial}{\partial\beta}R\right)^T. \tag{9}$$

If the weights w do not depend on the parameters, then

$$\nabla R\left(\begin{pmatrix}\theta\\ \tau\\ \widetilde{a}\\ \alpha\\ \beta\end{pmatrix}, w\right) = -2\begin{pmatrix}\sum_{i=1}^{n} w_i\left(Y_i - E\left(Y_i\right)\right)\sum_{l=1}^{n}\ln(h_i h_l)\dfrac{\left(h_i h_l\right)^{\theta}}{d_{i,l}{}^{\tau}}\exp\left(f_{(i,l)}\right)\\ -\sum_{i=1}^{n} w_i\left(Y_i - E\left(Y_i\right)\right)\sum_{l=1}^{n}\ln(d_{i,l})\dfrac{\left(h_i h_l\right)^{\theta}}{d_{i,l}{}^{\tau}}\exp\left(f_{(i,l)}\right)\\ \sum_{i=1}^{n} w_i\left(Y_i - E\left(Y_i\right)\right)\sum_{l=1}^{n}\dfrac{\left(h_i h_l\right)^{\theta}}{d_{i,l}{}^{\tau}}\exp\left(f_{(i,l)}\right)\\ \sum_{i=1}^{n} w_i\left(Y_i - E\left(Y_i\right)\right)\sum_{l=1}^{n}\dfrac{\left(h_i h_l\right)^{\theta}}{d_{i,l}{}^{\tau}}\exp\left(f_{(i,l)}\right)\left(c_{(i)} + c_{(l)}\right)^{T}\\ \sum_{i=1}^{n} w_i\left(Y_i - E\left(Y_i\right)\right)\sum_{l=1}^{n}\dfrac{\left(h_i h_l\right)^{\theta}}{d_{i,l}{}^{\tau}}\exp\left(f_{(i,l)}\right)g_{(i,l)}{}^{T}\end{pmatrix}, \tag{10}$$

where $f_{(i,l)} = \left(\widetilde{a} + \left(c_{(i)} + c_{(l)}\right)\alpha + g_{(i,l)}\beta\right)$.

In this case (w_i = const), the gradient method quickly gives estimates θ^*, τ^*, $\widetilde{a}^*$, α^*, β^*. Otherwise, the weights contain the unknown parameters. Therefore we must use an iterative procedure and successively recalculate estimates of the weights and the parameters.

4. An estimation of parameters *a* and σ^2

According to (3) and (7) we have the estimates for $i \neq l$:

$$E\left(Y_{i,l}\right)^* = \frac{\left(h_i h_l\right)^{\theta^*}}{\left(d_{i,l}\right)^{\tau^*}}\exp\left(\widetilde{a}^* + \left(c_{(i)} + c_{(l)}\right)\alpha^* + g_{(i,l)}\beta^*\right), \tag{11}$$

$$E\left(Y_i\right)^* = \sum_{\substack{l=1\\ i\neq l}}^{n} E\left(Y_{i,l}\right)^* = \sum_{\substack{l=1\\ i\neq l}}^{n}\frac{\left(h_i h_l\right)^{\theta^*}}{\left(d_{i,l}\right)^{\tau^*}}\exp\left(\widetilde{a}^* + \left(c_{(i)} + c_{(l)}\right)\alpha^* + g_{(i,l)}\beta^*\right). \tag{12}$$

Analogously from (6) we get

$$D\left(Y_i\right)^* = \left(\exp\left(\sigma^{2*}\right) - 1\right)\sum_{l=1}^{n}\left(E\left(Y_{i,l}\right)^*\right)^2. \tag{13}$$

At other hand we can estimate the variance of Y_i with respect to the variance definition

$$D\left(Y_i\right) = E\left(Y_i - E\left(Y_i\right)\right)^2.$$

Using $E\left(Y_i\right)^*$ as the estimate $E\left(Y_i\right)$, we have an alternative estimate of $D\left(Y_i\right)$:

$$D\left(Y_i\right)^{**} = \left(Y_i - E\left(Y_i\right)^*\right)^2. \tag{14}$$

Here we suppose a weak dependence between Y_i and $E\left(Y_i\right)^*$ because the last is calculated on base of many $\{Y_l\}$.

Now the variance parameter σ^2 can be estimated using the equalization of both values (13) and (14). By summing ones for $i = 1, \ldots, n$, we get

$$\sum_{i=1}^{n} D\left(Y_i\right)^* = \sum_{i=1}^{n} D\left(Y_i\right)^{**}$$

or

$$\left(\exp\left(\sigma^{2*}\right) - 1\right)\sum_{i=1}^{n}\sum_{\substack{l=1\\ l\neq i}}^{n}\left(E\left(Y_{i,l}\right)^*\right)^2 = \sum_{i=1}^{n}\left(Y_i - E\left(Y_i\right)^*\right)^2.$$

Therefore

$$\sigma^{2*} = \ln\left\{1 + \left(2\sum_{i=1}^{n-1}\sum_{l=i+2}^{n}\left(E\left(Y_{i,l}\right)^*\right)^2\right)^{-1}\sum_{i=1}^{n}\left(Y_i - E\left(Y_i\right)^*\right)^2\right\}. \quad (15)$$

Now the estimate of the parameter a is calculated as $a^* = \tilde{a}^* - \frac{1}{2}\sigma^{2*}$.

5. Balancing

Often one requests that the statistical data $\{Y_i\}$ and the estimates $\{Y_{i,l}^*\}$ have been balanced:

$$\sum_{l=1}^{n} Y_{i,l}^* = Y_i \ i = 1,...,n. \quad (16)$$

For that we introduce the correction coefficient $\delta_i > 0$ for each point i. Then corrected estimate of $Y_{i,l}$ is

$$\tilde{Y}_{i,l} = \delta_i Y_{i,l}^* \delta_l, \ \ i,l = 1,...,n. \quad (17)$$

To calculate the coefficients $\{\delta_i\}$ we have nonlinear system

$$\delta_i \sum_{l=1}^{n} Y_{i,l}^* \delta_l = Y_i, \ \ i = 1,...,n,$$

$$\delta_i = \left(\sum_{l=1}^{n} Y_{i,l}^* \delta_l\right)^{-1} Y_i, \ \ i = 1,...,n,$$

The experience shows that a solution is determined simply by the successive approaches method. For the k-th iteration

$$\delta_i^{(k)} = \left(\sum_{l=1}^{i-1} Y_{i,l}^* \delta_l^{(k)} + \sum_{l=i+1}^{n} Y_{i,l}^* \delta_l^{(k-1)}\right)^{-1} Y_i, \ \ i = 1,...,n, \quad (18)$$

where $\left\{\delta_i^{(0)}\right\}$ are initial values and we take in mind that $Y_{i,i}^* = 0$.

The iterations are ended, when a difference between two last values of $\delta^{(k)} = \left(\delta_1^{(k)} \quad \delta_2^{(k)} \quad ... \quad \delta_n^{(k)}\right)$ is less then prescribed precision $\varepsilon > 0$.

6. Numerical example

We use the suggested approach for the passenger railway transportations estimation between 23 member countries of the European Union. So here these countries play the part of the points.

Taking into account our previous investigation (Andronov, Zhukovskaya and Santalova, 2006) the following characteristics of the countries have been taken as regressors: c_1 is the average monthly labour cost, EUR; c_2, c_3, c_4 are gradation of countries upon intensity of use of air transport, of railway transport and of sea transport correspondingly; c_5 is a country gradation upon degree of popularity for tourism and c_6 is a country gradation upon the duration of membership in the EU.

As distances between points, the distances between capitals of countries have been taken. Y_i value is the railway transportations in thousands of passengers. The statistical data for the year 2007 have been obtained from the EuroStat database (EuroSTAT Yearbook, 2008). The values of gradation factors are determined by means of experts. For example, for Denmark we have $c_1 = 4.34$, $c_2 = 4$, $c_3 = 3$, $c_4 = 4$, $c_5 = 2$, $c_6 = 0$; for Latvia we have $c_1 = 0.68$, $c_2 = c_3 = c_4 = 1$, $c_5 = 0$, $c_6 = 1$.

Therefore it is necessary to estimate 15 parameters. The above described estimation procedure (for $w_i = 1$) gives the following values of estimated parameters:

$$\theta^* = 0.788, \ \tau^* = 2.786, \ \tilde{a}^* = -10.01,$$

$$\alpha^* = \left(0.074 \quad 0.061 \quad 0.077 \quad 0.063 \quad 0.078 \quad 0.33\right)^T,$$

$$\beta^* = \left(0.197 \quad 0.194 \quad 0.152 \quad 0.054 \quad 0.097 \quad 0.104\right)^T.$$

The corresponding value of criteria (8) R = 1.8×10^5. The multiple correlation coefficient (Srivastava, 2002) is equal to 0.97. The estimate σ^{2*} is calculated by formula (15): $\sigma^{2*} = 0.065$. Now using the estimate $\tilde{a}^* = -10.01$ we find

$$a^* = \tilde{a}^* - \frac{1}{2}\sigma^{2*} = -10.01 - \frac{1}{2}0.065 = -10.04.$$

Observed (Y_i) and estimated (Y_i^*) departures from each country in thousands of passengers are represented in the Table 1. The correction coefficients δ_i have been obtained as well.

Table 1. Estimation results

Country	Y_i	Y_i^*	δ_i	Country	Y_i	Y_i^*	δ_i
EU	44 270	41 210	-	Lithuania	7	0.99	7.062
Belgium	3 187	3 741	0.705	Luxembourg	2 333	905	2.882
Bulgaria	68	23	1.040	Hungary	631	164	3.424
Czech	1015	864	1.212	Netherlands	2 617	2 321	1.211
Denmark	4 974	5 557	0.806	Austria	1 575	2 017	0.362
Germany	5 072	5 279	0.956	Poland	353	202	1.487
Ireland	347	178	1.909	Portugal	98	41	1.629
Greece	18	6	2.050	Romania	197	37	3.888
Spain	329	164	1.554	Slovenia	97	37	2.064
France	5 184	5 240	0.948	Slovakia	1 459	1 081	2.939
Italy	1 600	890	1.654	Sweden	5 023	4 591	1.226
Latvia	2	0.85	2.354	UK	8 082	7 875	1.003

Analyzing the results, one can see that better estimates correspond to the old members of the EU. Probably, to improve the estimates for new members, it is necessary to consider ones separately.

8. Conclusion

The nonlinear regression model for the concrete correspondence $Y_{i,l}$ has been suggested. The unknown model parameters were estimated using the gradient method. Testing suggested the approach for passenger railway transportations estimation between member countries of the European Union show good results. Next we intend to continue our investigation and to improve the suggested model by extension of numbers of regressors and sample size.

References

Andronov, A. M.; Zhukovskaya, C. and Santalova, D. 2006. On Mathematical Models for Analysis and Forecasting of the Europe Union Countries Conveyances, in *Proceedings of the 46th RTU International Scientific Conference*, Riga Technical University. Riga, Latvia, 96–106.

EuroSTAT Yearbook. 2008. *The statistical guide to Europe. Data 1993-2007.* European Commission, EuroSTAT. Available from Internet: <http://epp.eurostat.ec.europa.eu>.

Sleeper, A. 2007. *Six Sigma Distribution Modeling*. McGraw-Hill, New York.

Srivastava, M. S. 2002. *Methods of Multivariate Statistics*. John Willey, New York.