The Third International Conference "Problems of Cybernetics and Informatics"
September 6-8, 2010, Baku, Azerbaijan. Section #4 "Applied Stochastic Analysis"
www.pci2010.science.az/4/27.pdf

PUBLIC TRANSPORT SERVICE QUALITY ESTIMATION ON THE BASIS STATISTICAL ANALYSIS

Irina Yatskiv[1], Nadezda Kolmakova[2], Vaira Gromule[3]

Transport and Telecommunication Institute, Riga, Latvia
"Rīgas Starptautiskā Autoosta" JSC, Riga, Latvia
[1]*ivl@tsi.lv*, [2]*nk@tsi.lv*, [3]*autoosta@autoosta.lv*

There is no unique approach to measuring service quality, but, it is accepted that the quality of service is usually a function of several particular quality attributes and determining of each factor weight is the important moment of measuring quality. Development of indicator of quality of services on the basis of regression model is considered in the paper. The numerical example of model estimation was done on the basis of results of questionnaire of transport experts about quality of service in Riga Coach Terminal. This enterprise provides the international, intercity and regional trips. The problems of the service quality provided by a Riga Coach Terminal have been considered by the paper's authors many times [1, 2]. The models presented in the papers had some assumptions, and in this paper we want to discuss the model without the assumption that the overall quality estimation and particular quality estimations are continuous variables. It is based on the approach which was offered in [3]. So, the dependent variable in this model is the overall quality estimation and the explanatory variables are the particular quality attributes. The aim – to estimate the unknown regression coefficients – the weights of particular quality attributes.

We describe the considered problem following [4]. The response Y_i of a concrete individual or item *i* (*i* = 1, …, *n*) is one of the fixed set of possible values, let $\{1, 2, \ldots, k\}$. These values are called ***categories.*** In our paper we suppose that the categories are ordered: the category *j* is "better" than *i* if *i* < ***j.*** The response probabilities $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ are function of vector of covariates or explanatory variables $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$ associated with the *i*-th individual: $\pi = \pi(x_i)$. We have at our disposal the matrix of the covariates *X* and a vector of responses *Y:*

$$X = \begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix}_{n \times d}, \quad Y = \begin{pmatrix} Y_1 \\ \ldots \\ Y_n \end{pmatrix}.$$

Our aim is to suggest the relationship between the response probability $\pi = \pi(x_i)$ and the explanatory variables $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$. For that aim we use an unobserved continuous random variable Z_i for the *i*-th individual:

$$Z_i = \sum_{j=1}^{d} \beta_j x_{i,j} + \zeta_i = x_i \beta + \zeta_i, \qquad (1)$$

where $\{\zeta_i\}$ are independent and identically normally distributed random variables with zero expectation and unknown variance $\sigma^2, \{\beta_j\}$ are unknown regression coefficients, $(\beta_1, \ldots, \beta_d)^T = \beta$.

Further, we introduce unknown parameters $\theta_1 < \theta_2 < \ldots < \theta_{k-1}$. If the unobserved variable Z_i lies in the interval $(\theta_{j-1}, \theta_j]$ then $y_i = j$ is recorded. Here *j* = 1, …, *k*, $\theta_0 = -\infty, \theta_k = \infty$.

We want to get maximal likelihood estimates of the unknown parameters $\beta = (\beta_1,...,\beta_d)^T$, $\theta = (\theta_1,...,\theta_{k-1})^T$ and σ. For that we have n observations with fixed values $\{Y_i\}$ and $x_i = (x_{i,1},..., x_{i,d})$.

Note that

$$\pi_j(x_i) = P\{Y_i = j\} = P\{\theta_{j-1} < Z_i \le \theta_j\} == \Phi\left(\frac{\theta_j - x_i\beta}{\sigma}\right) - \Phi\left(\frac{\theta_{j-1} - x_i\beta}{\sigma}\right) \tag{2}$$

We see that the unknown parameters $\beta = (\beta_1,...,\beta_d)$, $\theta = (\theta_1,...,\theta_{k-1})$ and σ can be estimated up to constant factor. Therefore we use united parameters $\widetilde{\beta} = (\widetilde{\beta}_1,...,\widetilde{\beta}_d) = \beta / \sigma$ and $\widetilde{\theta} = (\widetilde{\theta}_1,...,\widetilde{\theta}_{k-1}) = \theta / \sigma$. Note that situation takes place often in the econometrics [5, 6].

To rewrite down the corresponding log-likelihood function, we rearrange our observation (individuals) as follows: at first observations of the first category are written, then the second category and so on. Let n_j be an observation size for the j-th category,

$$n_0 = 0,\ n_1 + ... + n_k = n,\ N_0 = 0,$$
$$N_j = N_{j-1} + n_j,\ \ j = 1,\ ..., k-1;\ N_k = n$$

Then the log-likelihood function can be written as

$$l(\widetilde{\beta}, \widetilde{\theta}; y) = \sum_{j=1}^{k} \sum_{i=N_{j-1}+1}^{N_j} \log\left(\Phi(\widetilde{\theta}_j - \widetilde{\beta}x_i) - \Phi(\widetilde{\theta}_{j-1} - \widetilde{\beta}x_i)\right) \tag{3}$$

where Φ is the cumulative standard normal distribution function, $\widetilde{\theta}_k = \infty$.

Following the usual technique, let us derive the derivatives of the log-likelihood function with respect to $\widetilde{\theta} = (\widetilde{\theta}_1,...,\widetilde{\theta}_{k-1})$ and $\widetilde{\beta} = (\widetilde{\beta}_1,...,\widetilde{\beta}_d)$.

The derivatives of the log-likelihood function with respect to $\widetilde{\theta}_j$ for $j = 1, \ldots, k-1$ are

$$\begin{aligned} \frac{\partial l}{\partial \widetilde{\theta}_j} = & \sum_{i=N_{j-1}+1}^{N_j} \left(\Phi(\widetilde{\theta}_j - \widetilde{\beta}x_i) - \Phi(\widetilde{\theta}_{j-1} - \widetilde{\beta}x_i)\right)^{-1} \times \\ & \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(\widetilde{\theta}_j - \widetilde{\beta}x_i)^2\right) - \sum_{i=N_j+1}^{N_{j+1}} .\left(\Phi(\widetilde{\theta}_{j+1} - \widetilde{\beta}x_i) - \Phi(\widetilde{\theta}_j - \widetilde{\beta}x_i)\right)^{-1} \times \\ & \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(\widetilde{\theta}_j - \widetilde{\beta}x_i)^2\right). \end{aligned} \tag{4}$$

The derivatives of the log-likelihood function with respect to $\widetilde{\beta} = (\widetilde{\beta}_1,...,\widetilde{\beta}_d)^T$:

$$\begin{aligned} \frac{\partial l}{\partial \widetilde{\beta}} = & -\sum_{j=1}^{k} \sum_{i=N_{j-1}+1}^{N_j} \left(\Phi(\widetilde{\theta}_j - \widetilde{\beta}x_i) - \Phi(\widetilde{\theta}_{j-1} - \widetilde{\beta}x_i)\right)^{-1} \times \\ & \times \frac{1}{\sqrt{2\pi}\sigma} \left\{\exp\left(-\frac{1}{2}(\widetilde{\theta}_j - \widetilde{\beta}x_i)^2\right) - \exp\left(-\frac{1}{2}(\widetilde{\theta}_{j-1} - \widetilde{\beta}x_i)^2\right)\right\} x_i \end{aligned} \tag{5}$$

.

Now we should solve the maximal likelihood equations

$$\frac{\partial l}{\partial \tilde{\beta}} = 0; \ \frac{\partial l}{\partial \tilde{\theta}_j} = 0, \quad j = 1,...,k-1. \tag{6}$$

Our experiences show that a solution is gotten easily by using standard computer programs. In other case it is necessary using Taylor expansion for a solution of the maximal likelihood equations (6).

If parameter estimates $\tilde{\theta}^* = \left(\tilde{\theta}_1^*,...,\tilde{\theta}_{k-1}^*\right)^T$ and $\tilde{\beta}^* = \left(\tilde{\beta}_1^*,...,\tilde{\beta}_d^*\right)^T$ are known, it is possible to estimate the probability of interest (2):

$$\pi^*_j(x_i) = P^*\{Y_i = j\} = P\{\tilde{\theta}^*_{j-1} < Z_i \le \tilde{\theta}^*_j\} = \Phi\left(\tilde{\theta}^*_j - x_i\tilde{\beta}^*\right) - \Phi\left(\tilde{\theta}^*_{j-1} - x_i\tilde{\beta}^*\right). \tag{7}$$

Numerical Example.

In our research we used for model estimation the homogeneous sample of transport experts. 100 hundred questionnaires were distributed to national transport experts in spring 2009. A total of 44 questionnaires from experts were returned. The questionnaire included 22 particular attributes of quality distributed on groups: accessibility (availability); information; time characteristics of service; customer service; comfort; safety; infrastructure and environment. Also the overall quality of service was evaluated. As well as particular attributes of quality the overall quality service was estimated on a scale 0-5.

We consider the next method of forming the independent variables for regression model, in particular: we will form 7 new variables corresponding to 7 groups of attributes (categories of questions). Grouping of the initial attributes and calculation of new values on the basis of a arithmetic mean leads to a replacement of categorical variable x_i (i=1,..,22) by interval w_l (l=1,..,7). In Table 1 the descriptive characteristics for new "grouped" variables are presented.

Table 1. Descriptive measures of the grouped variables (attributes of quality) and overall quality

	Mean	Median	Mode	Frequency	Min	Max	Std.Dev.
W1	4.341	4.500	5.000	22	1.000	5.000	0.861
W2	4.064	4.000	4.670	13	2.000	5.000	0.678
W3	4.205	4.250	4.000	14	3.000	5.000	0.442
W4	4.031	4.170	Multiple		2.500	4.830	0.620
W5	3.557	4.000	4.000	18	1.000	5.000	0.916
W6	3.545	3.750	4.000	12	1.500	5.000	0.901
W7	3.119	3.000	3.000	12	1.000	5.000	0.845
W8	3.931	4.000	4.000	28	3.000	5.000	0.528

Let's use method described above.

Build-in optimization block in Mathcad (functions *Given* and *Find*) was used to find optimal values of estimated parameters $\beta = (\beta_1,...,\beta_5)^T$, $\theta = (\theta_1,...,\theta_4)^T$. In program we set $\theta_0 = -3,\ \theta_5 = 8.$ The initial values of the parameters are

$$\beta = (0.01 \ \ 0.01 \ \ 0.01 \ \ 0.01 \ \ 0.01 \ \ 0.01 \ \ 0.01)^T \text{ and } \theta = (1 \ \ 2 \ \ 3 \ \ 4)^T.$$

After optimization part we get the following estimations of unknown parameters:

$\beta = (1.18 \ \ -0.639 \ \ -0.719 \ \ -0.48 \ \ 0.522 \ \ 1.002 \ \ 0.51)^T$ and $\theta = (-0.186 \ \ 0.735 \ \ 3.015 \ \ 6.539)^T$.

The optimum value of likelihood function is $l\left(\tilde{\beta}^*, \tilde{\theta}^*\right) = -20.384$. Checking gives correct results because all partial derivatives are equal to zero.

Table 2 contains values of expert estimates and predicted estimates according to formula (7).

Table 2. Real and predicted estimates

Expert Number, *i*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*
Expert estimates	1	**2**	**3**	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4
Predicted estimates	1	**3**	**4**	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4
Expert Number, *i*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*	*39*	*40*	*41*	*42*
Expert estimates	4	4	4	4	4	4	4	4	4	4	4	4	**4**	4	4	4	4	**5**	**5**	**5**	**5**
Predicted estimates	4	4	4	4	4	4	4	4	4	4	4	4	**5**	4	4	4	4	**4**	**4**	**4**	**4**

As we can see the predicted values are close to real response vector *Y*. The considerable quantity of deviations is observed in categories with small amount of observations. It can be connected with sample imperfection.

Our example shows that developed model for an integrated quality indicator allows not only to compare and reveal the significant categories of qualities influencing the general indicator, but also to predict customer choice. The model constructed allows comparing analyzed service to the services given by other companies and to simplify monitoring of quality indicators.

References

1. Yatskiv I., Gromule V., Medvedevs A. Development the System of Quality Indicators as Analytical Part of the Information System for Riga Coach Terminal. In: *The International Conference "Modelling of Business, Industrial and Transport Systems, RTU, Riga, 2008,* pp. 278-283.
2. Yatskiv I., Gromule V., Kolmakova N., Pticina I. Development of the Indicator of Service Quality at Riga Coach Terminal. In: *Proceeding of the 9th International Conference Reliability and Statistics in Transportation and Communication,* Transport and Telecommunication institute, Riga, 2009, pp. 124-133.
3. Alexander Andronov, Nadezda Kolmakova, Irina Yatskiv. A regression model for polytomous data and its application. In: *Proceeding of the 11th WSEAS International Conference on MATHEMATICS AND COMPUTERS IN BUSINESS AND ECONOMICS (MCBE '10)*, G. Enescu University, Iasi, 2010.
4. McCullagh P. and Nelder J.A. *Generalized Linear Models. 2nd ed.* Chapman & Hall/CRC, 1989.
5. Andronov A. Maximal Likelihood Estimates for Modified Gravitation Model by Aggregated Data. In: *Proceedings of the 6th St. Petersburg Workshop on Simulation*, St. Petersburg State University, St. Petersburg, 2009, pp. 1016-1021.
6. Andronov A., Santalova D. On Nonlinear Regression Model for Correspondence Matrix of Transport Network. *Selected papers of the International Conference Applied Stochastic Models and Data Analysis* (ASMDA-2009). L.Sakalauskas, C.Skiadas and E.K.Zavadskas (Eds.), Vilnius Technical University, Vilnius, 2009, pp. 90-94.