Market for statisticians in developing economies: The case study of Pakistan's corporate sector

Javed Iqbal and Abbas Rawish

Department of Statistics University of Karachi and Monash University

January 2007

Market for statisticians in developing economies: The case study of Pakistan's corporate sector

Javed Iqbal[1] and M. Abbas Rawish[2]

Abstract

Statisticians are the professionals responsible for collection, presentation and analysis of data and help decision making in face of uncertainty. They are the backbone of any institution or activity facing decision under risk and uncertainty. Statistics provides a cost-effective solution for analyzing the data by making decision for a larger set of data based on information from only a smaller part of it. The professional services of statisticians are required in all fields of life for analyzing data and interpretation of the resulting output. Corporate sector is also an important employer of the services of statisticians. This paper is aimed at exploring the factors affecting market for statisticians in the corporate sector of a large urban centre of a developing country. An econometric analysis is carried out using a count data model to explain the factors responsible for the firms employing the services of statisticians. The Count Data models have found several applications in statistics, health economics, recreational studies and finance. This paper contributes to the literature by providing yet another interesting application of the count data model by examining the market for professionals.

KEY WORDS: Statisticians, Count Data Modes, Poisson Regression, Negative Binomial Regression, Developing Economy

[1] Corresponding author: Department of Econometrics and Business Statistics Monash University and Department of Statistics University of Karachi
Postal Address: Unit 2, 74 Ormond Road, Clayton 3168, Victoria, Australia.
Email: Javed.Iqbal@Buseco.monash.edu.au

[2] Department of Statistics Federal Urdu University of Arts, Sciences & Technology, Karachi

Introduction

Statisticians are the professionals responsible for collection, presentation and analysis of data and help management of organizations in making decisions in face of uncertainty. They are the backbone of any institution or activity facing decision under risk and uncertainty. The uncertainty comes partly due to the fact that usually a part of entire aggregate or population is examined and the results drawn from it are generalized for whole population. Thus statistics provides a cost-effective solution for analyzing the data and making decision for a larger set of data based on information from only a smaller part of it. Statistics has found applications in a wide variety of disciplines such in business and finance, engineering, medical and biological sciences and earth sciences. The professional services of statisticians are required in all of these fields for analyzing data and interpretation of the resulting output. Corporate sector is also an important employer of the services of statisticians.

This paper is aimed at exploring the determinants of market for statisticians in the corporate sector of a large urban centre of a developing country. The market specifically considered is the corporate sector in Karachi which is Pakistan's largest city and an important centre of business, trade and industry. It is also home to the largest and most active of the three stock exchanges in Pakistan. More than 700 firms are listed in the Karachi Stock Exchange. Approximately a quarter of these firms have their head offices or area of operation in Karachi. It is of interest to examine the factors or determinants of the demand for statisticians in the corporate sector of the developing country. An econometric analysis is carried out using count data models to explain the number of

statisticians employed by the corporate firm. As the dependent variable is discrete the relevant econometric model to be employed for examining these determinates is the count data model based on the Poisson and Negative Binomial regressions. Count data models have found several applications in health economics, recreational studies and finance. This paper contributes to the literature by providing yet another interesting application of the count data model by examining the market for professionals.

An important potential determinant for the response variable is the level of economic activity of the firm. The firms with a higher level of manufacturing, sales and export are more likely to use the services of the professionals of the fields than firms with a lower level of activity. The level of business activity is captured by the gross sales of the firm. The firms operating in some industrial sector are likely to generate more data to be analyzed than firms in other sectors. The specification of the sectors of the firms is indicated by sectoral dummy variables. The sectors considered are Cement, Sugar, Fuel and Energy, Textile, Chemical, Engineering, Paper and Board and other miscellaneous sector. Internationally recognized quality standard such as ISO 9000 and later versions are becoming increasingly important for the firms in developing countries to sell their products to local and international markets. Professionals having knowledge of Statistical Quality Control are of great benefit for to these firms. A dummy variable is therefore employed to indicate whether or not the firm has a quality control section/departmement. Market research is also one of the fields specially benefiting form the services of the statisticians therefore a dummy variable to indicate the presence of a market research department is employed. A similar dummy variable is employed if the firms have a research and development section/department. Two further dummy variables are

employed to investigate whether it makes difference with respect to demand for statistician if the firm operates in public rather than private sector and whether it is a multinational or a local firm.

This paper contributes to the literature in several ways. The studies aimed at investigating characteristics for the labor market of professionals are conducted mostly for the developed countries. Very little is known regarding the developing countries in this regard especially for the market for statisticians. Further, the demand for goods and services are well studied in economics and econometric literature but perhaps few studies, if none, exist aimed at examining the demand for professional services of statisticians. Such studies are important for the potential graduates of the respective fields as well as the educational institutions which can set a better courses structure reflecting the requirement of the market for these professionals. Besides investigating the market for the statistics professionals, this paper provides another interesting application of the count data model by studying the determinants of the market for statisticians.

Survey Design and Data Sources

The target population consists of all corporate sector firms listed at Karachi Stock Exchange (KSE) with head offices or area of operation in Karachi. The firms listed at KSE have a varied type of activities and operations, which are categorized in textile, fuel and energy, cement, paper and board, sugar, chemical, engineering, transport and communication and others sectors. As the population of firms is quite heterogeneous with respect to their size (measured by fixed assets) and level of business activity (measured by sales), a stratified random sample of firms was selected with three strata i.e. small, medium and large firm formed according to the gross sales. The boundaries of the

sales categories were arbitrarily set as 500, 1000 and 2000 million rupees respectively. Cochran (1986) also advocates stratified random sampling of firms of different sizes.

The primary data on the number of statistician working in the firm are collected from a survey conduced in mid of 2003 by personal interview method. A questionnaire was developed with questions on number of statisticians working and other question related to this variable of interest. An important consideration was what defines a 'statistician'. One approach is to relate it with the person having a honors or graduate degree in Statistics. Shettle and Gaddy (1998) have defined Mathematics and Probability, Biometrics and Biostatistics, Econometrics, Psychometrics and Social Statistics as Statistics field. It is, however, generally true that the most important users of Statistics are non-statisticians. Also persons having related degree can perform similar tasks as that of statisticians. It was also considered that the respondents in a developing economy's firms might not be fully aware of activities that a statistician does or can perform. Therefore a list of statistical activities was provided in the questionnaire. The list includes collecting, presenting and analyzing data, planning a controlled experiment to identify important factors affecting the output, forecasting sales or demand, building mathematical models to explain some variable of interest with some microeconomic or macroeconomic variables, performing statistical quality control, summarizing information from a large group of related variables into one or more index, determining the number of units of a product to be produced to maximized profit subject to resource constraints, studying the demographic or social characteristics of population, constructing life tables and determining premium for insurance. An employee of the firm engaged in these activites was defined as 'statistician' for our purpose. It was desired to know the characteristics of

the firms employing statistician. The questions about presence of quality control, market research and R&D departments or sections within the firm were therefore included.

The respondents of the survey were Admin Manager, Human Resource Manager, Personal Relation Officer or Statistical Officer. The primary source for the list of the firms is the document published annually by the State Bank of Pakistan titled 'Balance Sheet Analysis of Joint Stock Companies listed under Karachi Stock Exchange' which contains the list of the firms, their addresses and contact details. It was observed that in the year 2003, 257 firms have the head offices or area of operation in Karachi. As the modeling and inference techniques of count data models are based on large sample theory it was decided to select a larger sample of about 50 % of the firms. Excluding non-responses and incomplete questionnaires finally a sample of 101 firms is selected for analysis. The data on sales, sectoral belonging and public/private classification is also obtained from the SBP document. The classification of multinational/local is determined by common knowledge and from web resources. Industrial Almanac of Pakistan (2000) is another source of some of the similar information.

Models of Count Data

In many business and economics studies the variable of interest is discrete that can be modeled by some of the many discrete probability models. The econometrics analyses of count data focus on linking the response count variable to a number of covariates that could possible explain the variation in former. Some examples include number of visits to a doctor linked to some economic and demographic variables, number of airline accidents explained by variables on profitability and size of the airlines, number of takeover bids received by a target firm explained by bid premium, firm size and defensive action by

firms. Cameron and Trivedi (1996, 1998) present several other examples in economics and finance. This paper adds another application in this literature by studying the labor market for statisticians in the corporate sector of a developing economy. Our interest center on whether the market can be explained by firm's level of economic activity, the industrial sector and organizational structure.

The Poisson regression model has been widely used to study the data of count nature. This model assumes that the discrete count variable y_i conditional of on the covariates X_i is given by

$$f(y_i \mid X_i) = \frac{e^{\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots \tag{1}$$

To ensure non-negativity the conditional mean is parameterized in the log-linear form

$$\lambda_i = \exp(X_i'\beta) \tag{2}$$

This is essentially a non-linear regression model. Given the data on response y_i and covariates X_i the Maximum Likelihood is the most efficient estimation method for the model parameters. The Poisson regression model implicitly assumes the equality of conditional mean and variance i.e.

$$E(y_i \mid X_i) = V(y_i \mid X_i) = \exp(X_i'\beta) \tag{3}$$

Most economics and financial data are overdispersed that is the conditional variance exceeds the mean. Agresti (2002) elaborates that a common cause of overdispersion is heterogeneity of the sampling units. In the context of our application suppose mean number of statisticians employed by a firm (y) is affected by sales of the firm, sector to which it belongs and structure of the firm and suppose y has a Poisson distribution at each fixed combination of these predictors. The firms with a specific sales value may vary due to different sectors belonging and difference in their structure. Thus population

of firms with a certain sales is a mixture of several Poisson populations each having its own means for the response. This heterogeneity results in greater variation at this sales value than predicted by the Poisson model. For ordinary regression models with normal distribution of y overdispersion is not an issue because this distribution has a separate parameter (the variance) to describe the variability. Poisson distributions however the variance is a function of mean. The alternative Negative Binomial regression model overcomes both of these shortcomings. The Negative Binomial model with conditional mean and variance λ_i and $\lambda_i + \alpha\lambda_i^2$ respectively is given by

$$f(y_i \mid X_i) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(\alpha^{-1})\,\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\lambda_i + \alpha^{-1}} \right)^{\alpha^{-1}} \left(1 - \frac{\alpha^{-1}}{\lambda_i + \alpha^{-1}} \right)^{y_i} \quad \text{y}_\text{i} = 0, 1, 2, \ldots \qquad (4)$$

The index α is called dispersion parameter. As $\alpha \to 0$, $Var(y_i) \to \lambda_i$ and Negative Binomial distribution converges to the Poisson distribution. Usually α is unknown. Estimating it helps summarize the extent of overdispersion. The conditional mean in this case can also be expressed as in (2) and the Maximum Likelihood method can be employed to estimate the model parameters.

Empirical Analysis

Table 1 presents the frequency distribution of the number of statisticians employed by the sample firms. Approximately 31 % firms do not employ any statistician. Sample mean and sample variance of the count are 6.554 and 137.00 respectively. The probability of zero count from the Poisson distribution with the estimated mean as the parameter is 0.1438 which is much lower than the observed relative frequency of the counts. For the Negative Binomial distribution the dispersion parameters is estimated to be 0.329.

Employing this estimate the estimated probability of the zero count is 0.367 which is much closer to the observed relative frequency. It appears that at least for zero count the Negative Binomial model can better represents the observed data. The formal tests of overdispersion which can also be considered as the nested tests of the Poisson versus the Negative Binomial model however need to be performed before proceeding further.

Several such tests are discussed in the literature. Cameron and Trivedi (1986) devised an auxiliary regression based test of the following form:

$$\left[\frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i}\right] = \alpha\hat{\lambda}_i + \varepsilon_i$$

The t-statistic of the estimated slope $\hat{\alpha}$ is asymptotically normal under the null hypothesis of no overdispersion against the overdispersion of Negative Binomial form. The estimated regression equation is

$$\left[\frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\lambda_i}\right] = 0.3326\ \hat{\lambda}_i$$

The t-value of the coefficient is 5.235 with a p-value nearly zero thus rejecting equidispersion hypothesis of the Poisson form. Cameron and Trivedi (1996) also propose a Likelihood Ratio test for the nested hypothesis of overdispersion which is given as

$$\text{LR} = 2(\text{ULLF-RLLF}) \quad \sim \chi^2(1)$$

Here ULLF is fitted unrestricted maximum log likelihood value for the Negative Binomial model and RLLF is restricted maximum log likelihood value for Poisson model. Application to our count data gives LR = 2(-208.2443+281.1845) = 145.88 with a p-value near zero. Once again the Poisson restriction is rejected in favor of the overdispersed Negative Binomial alternative.

Table 2 presents the Maximum Likelihood estimation result for the conditional mean of the following form for both Poisson and Negative Binomial models.

$$\lambda_i = \exp(\beta_0 + \sum_{j=1}^{13} \beta_j X_{ij}) \quad (5)$$

The t-values are based on the robust Huber/White standard errors. The Sugar sector represents the reference category for the sectoral dummies. This sector is the least likely to have a statistician as observed in our sample. Judging from the three model evaluation criteria the Negative Binomial regression model appears to be superior to the Poisson model. This is also consistent with the results of the overdispersion test where the Poisson model is rejected in favor of the negative Binomial alternative. The later model also results in lower standard errors and therefore higher t-values. The log likelihood for the Negative Binomial model is much higher, the Akaike Information Criteria is lower and the Pseudo R^2 is better than the Poisson model. The Pseudo R^2 (also called the likelihood ratio index) is defined as one minus the ratio of fitted maximum log likelihood to the log likelihood value of the model that only includes a constant. The lowest possible likelihood value is that of the model with only a constant. This criterion measures the actual log likelihood gain as a ratio of the potential log-likelihood gain. In both models the sector dummies are not significant at 5 % level of significance. The dummy variables for the ownership status of Private versus Public or multinational versus local firms are also not significant.

It is of interest to test some joint hypothesis also regarding the factors that might significantly explain the number of statistician employed by the firm. Specifically we would like to know does industrial sector as a whole in which the firm is classified matter. Does the ownership structure matter? i.e. does it make difference if the firm is

privately owned rather than a public entity or that the firm is a multinational rather than of a local origin. It is also of interest to investigate whether the organizational structure within the firm i.e. presence of some special department/section within the firm such as market research, research and development and quality control can significantly explain the decision by the firms to employ the services of statisticians. These hypotheses can be easily tested by imposing the joint zero restrictions on the estimated model. The Likelihood Ratio test can be employed which is asymptotically distributed as Chi-Square with number of degrees of freedom equal the number of restrictions to be tested. In table 3 the results of tests are presented for the Negative Binomial case. The joint tests also corroborate the finding form the individual tests. The variables regarding the organizational structure within the firm are important in that for the selected model the coefficients for this variable are significant in individual as well as in the joint hypothesis tests.

Table 4 reports the more parsimonious Negative Binomial regression ignoring the dummy variable for sector and ownership structure. This specification slightly improves the Akaike Information Criteria but slightly deteriorates the fit of the model as measured by the Pseudo R^2 and the log likelihood value. The coefficients have their expected signs. The interpretation of a coefficient is that a one unit increase in explanatory variable X_j increases the number of statistician employed by the firm by $\beta_j \times 100$ percent. This implies that an increase of 309 million rupees in sales increase the number of statistician employed by 1 %. The number of statistician is on the average 230 % (2.3 times) higher in the firm that has a quality control department compared to the firm which does not

have. Similarly the number of statistician employed is 1.3 times higher if the firm has an R & D department compared to when it does not have. A firm which has market research activity has on the average 43 % more statistician than otherwise.

Conclusion

This paper applies the count data model to investigate the labor market for statisticians in the corporate sector of a developing economy. It is found that the level of business activity as measured by sales and the organizational structure within the firm are important determinants of number of statisticians a firm employs. The firms with a higher level of economic activity are more likely to require professional's services for the data analysts and their interpretation. The firms that accept the competitive challenge of local or export market pay special attention to quality control and research and development. The statistician may be of great value to these firms. They assist in ensuring the quality of the products and help designing the improved products. The market research activity possesses considerable scope for survey design, analysis and interpretation for the data as indicated by their employment of professional for these services. The industrial or ownership structure of the firms does not appear to be relevant in explaining the employment of statisticians. All theses are useful information for the graduates entering the job market in statistics and related fields. The educational institutions preparing the graduates for the market also need to focus attention to specialize some of their courses to fulfill the need for the corporate sector in the developing economies.

ACKNOWLEDGEMENT: The authors would like to thank Junaid Sagheer Siddiqui, Asim Jamal Siddiqui and Ehtesham Hussain for their helpful comments on the

questionnaire. The first author thankfully acknowledges the research grant from the office of Faculty of Science University of Karachi.

Reference

Agresti A. (2002) Categorical Data Analysis (New York: John Wiley and Sons).

Balance Sheet Analysis of Joint Stock Companies (2003) State Bank of Pakistan.

Cameron, A.C. and Trivedi, P.K. (1986) Econometric models based on count data: Comparisons and applications of some estimators and tests, *Journal of Applied Econometrics*, 1, pp. 29-54.

Cameron, A.C. and Trivedi. P.K. (1996) Count data models for financial data, *Handbook of Statistics*, 14, pp. 363-391

Cameron, A.C. and Trivedi. P.K. (1998) Regression analysis of count data, (Cambridge University Press).

Cochran, G.W. (1986) Sampling Techniques, (New York: John Wiley).

Industrial Almanac of Pakistan (2000) (Karachi: Vital Information Services).

Shettle C. and Gaddy, C. (1998) The labor market for statisticians and other scientists, *The American Statisticians*, 52, pp. 295-302.

Table 1: Frequency distribution of the number of statistician in the sample firms

Count	0	1	2	3	4	5	6	7	8	9	≥ 10
Frequency	31	12	9	8	5	3	7	3	0	2	21
Relative Frequency	0.307	0.119	0.089	0.079	0.050	0.030	0.069	0.030	0	0.019	0.208

Table 2: Poisson and NB regression parameter estimates and t ratios based on Huber/White robust standard errors

Variable	Poisson		Negative Binomial	
	Coefficient	t-Stats	Coefficient	t-Stats
Constant	-0.958	-1.463	-0.970	-1.457
Sales	3.24×10^{-5}	2.782	2.10×10^{-5}	3.324
Cement	-0.144	-0.185	0.127	0.169
Engineering	0.393	0.544	0.621	0.898
Chemical	0.443	0.584	0.755	1.058
Fuel & Energy	-1.382	-1.016	-0.405	-0.496
Paper & Board	0.847	1.137	1.016	1.393
Textile & Related	-0.109	-0.150	0.112	0.160
Other Industries	0.347	0.442	0.680	0.968
Private	-0.325	-1.196	-0.303	-1.031
Multinational	0.267	0.868	0.214	0.841
Quality Control	2.525	6.263	2.142	6.388
R& D	0.880	3.870	0.849	4.010
Market Research	0.271	1.143	0.543	2.891
Log Likelihood	-281.185		-208.244	
AIC	5.845		4.420	
Pseudo R^2	0.645		0.737	

Table 3: Testing some joint hypothesis of interest with the Likelihood Ratio test. The p-values are from the Chi Square distribution with the number of degrees of freedom equal the number of parametric restrictions.

	Hypothesis	Likelihood Ratio	P-value
Industrial Sector	$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$	10.104	0.1827
Ownership Structure	$H_0 : \beta_9 = \beta_{10} = 0$	1.082	0.5821
Organizational Structure	$H_0 : \beta_{11} = \beta_{12} = \beta_{13} = 0$	88.135	0.0000

Table 4: The Parsimonious Negative Binomial Regression

Variables	Coefficients	t-statistics	P-values
Constant	-0.984	-3.376	0.001
Sales	3.24×10^{-5}	2.984	0.003
Quality Control	2.297	6.878	0.000
R& D	1.335	6.505	0.000
Market Research	0.425	2.161	0.031
Log Likelihood		-215.458	
AIC		4.385	
Pseudo R^2		0.728	