Modeling of Job Tenure: Insights from Russia

Fengchen Wang*, William Attatsitsey, Romie F. Littrell, Natalia Volkova
*fwang@hse.ru, chiefwilliams8@gmail.com, rlittrell@hse.ru, nv.volkova@hse.ru

School of Economics and Management, National Research University-Higher School of Economics (HSE University), St. Petersburg, Russia

Abstract; Over the years, both business practitioners and social scientists have been concerned about employee turnover. Several attempts to estimate the job tenure of an individual given specific criteria have been made as a result of this. An index called "job tenure" shows how stable a person's employment is over time. One measure of loyalty in the workplace is length of employment. Employee pleasure is reflected in loyalty, which raises productivity and, in turn, increases business profitability. With the aid of the categorical regression model with optimal scaling technique (CATREG) and CV data from HeadHunter, the largest job board in Russia, this study uses data from Russia and takes into account the employee's age, gender, and educational levels to build a model that anticipates their employment tenure. Our findings make it abundantly evident that, in the case of the Russian labor market, the older the job seeker or an employee is and the better educational level they possess, the longer employment duration may be anticipated within an organization.

Keywords: Job tenure, employee retention, categorical regression model (CATREG), demographic variables, people analytics, digital HRM, job board, Russia

1 INTRODUCTION

The length of an employee's employment, as defined by [1], is their job tenure. Employee affective commitment at work was found to be correlated with tenure. One measure of employee loyalty in the workplace has been job tenure. Employee satisfaction, productivity, and business profitability are all moderated by loyalty. As academics, we contend that a worker's educational background, generation (age group), and gender have an impact on the boredom and lack of motivation that come with holding a position for a longer period of time. Researchers have been examining the relationship between financial and non-financial factors that affect employee job tenure, or employment turnover, using a variety of indicators, including salaries, social benefits, and vocational training as financial factors, and organizational structure of position and status, work-life balance, and relationships among coworkers as non-financial factors.

When examining developing and emerging economies, previous studies did not place a lot of emphasis on demographics like age, gender, and education. Utilizing data gathered locally, these investigations were carried out in the setting of industrialized economies [2]-[4]. We aim to address this perceived gap in the research on emerging economies like Russia. Additionally, in Russia, the cost of turnover is at least three times the employee's pay. Costs related to voluntary departures range from 25% to 500% of an employee's gross salary [5]. In light of this, the goal of this study, therefore, is to further be understanding on turnover by

ICICA 2022, December 02-04, Chongqing, People's Republic of China

DOI 10.4108/eai.2-12-2022.2327918

investigating the impact of demographic factors including age, gender, and educational levels on the length of employment tenure of job seekers.

2 RESEARCH METHODOLOGY

2.1 Sample data and descriptive statistics

The study used a secondary data source that includes anonymized CVs of job seekers from the Saint Petersburg area that were posted by job seekers as of August 2019 to the massive Russian job portal — HeadHunter (hh.ru). In accordance with the research goal, the original data were cleaned and reorganized to provide a truthful total of 18,664 CVs, whose descriptive statistics are shown in Table 1.

The sample dataset consisted of over 21 job specialization fields for job seekers working or worked in the tertiary sectors; a summary of these specialization fields is presented in Table 2. There were 18,664 CVs with the specialty "IT, Internet and Telecom industries," which made up 39.0% of the sample dataset. The percentage of job seekers with the specializations "financial services, and professional services" and "marketing, advertising, PR" is 6.8% and 6.2%, respectively, and are ranked second and third overall in the sample data.

Table 1. Descriptive statistics of sample data

Variables	**N**	**%**
Gender		
Female	11404	61.1
Male	7260	38.9
Generation (Age Group)		
Gen X (born 1960~1979)	3691	19.8
Gen Y (born 1980~1994)	14399	77.1
Gen Z (born 1995~2010)	574	3.1
Educational Level		
Secondary education	1369	7.3
Bachelor’s degree	1253	6.7
Master’s degree (or equivalent specialist diploma)	15546	83.3
Ph.D. or equivalent doctoral degree	496	2.7
Work Experience (by Year)		
< 0.5 years	1114	6.0
0.5 years – 1 year	2693	14.4
1year – 2 years	7233	38.8
2 years – 3 years	4217	22.6
3 years - 5 years	2509	13.4
5 years – 10 years	776	4.2
10 years – 15 years	90	0.5
>15 years	32	0.2
Total	18664	100.0

Table 2. Distribution of frequencies by job specialization

Job Specialization	**N**	**%**
IT, Internet, Telecom	7281	39.0%
Financial services, and professional services	1266	6.8%
Marketing, advertising, PR	1155	6.2%
Administrative staff	1133	6.1%
Production	1038	5.6%
Executive management	1014	5.4%
Tourism, hotels, catering	949	5.1%
Sales	881	4.7%
Transport, logistics	659	3.5%
Science, education	632	3.4%
Construction, real estate	574	3.1%
HR management, trainings	509	2.7%
Health services	329	1.8%
Lawyers	271	1.5%
Procurement	270	1.4%
Art, entertainment, media	250	1.3%
Extraction of raw materials	196	1.1%
Automobile business	95	0.5%
Installation and service	63	0.3%
Public service, non-profit	40	0.2%
Security	31	0.2%
Others	28	0.2%
Total	18664	100

2.2 Statistical method and equation

One of the most popular statistical methods in social science research that may be used to evaluate data from causal-comparative, correlational, or experimental investigations is the linear regression model [6]. Univariate regression analysis, also known as simple linear regression, refers to the regression that uses only one independent (predictor) variable, whereas multivariate regression analysis, also known as multiple linear regression, refers to the regression that uses multiple independent (predictor) variables [7], [8].

The decision to use standard linear regressions or multiple regressions is heavily influenced by the types of predictors and outcome variables since continuous or nearly continuous variables are the best fit for standard multiple regression. Although measurement is recorded on scales with an ambiguous unit of measurement, research in social and behavioral science frequently yields categorical or qualitative data that are non-numerical [9]. Even if categorical predictors can be transformed using specialized statistical coding techniques like dummy coding, effects coding, orthogonal coding, or criterion coding to fit the standard regression model, the transformation process for categorical predictors will be very difficult when there are multiple variables with a large number of levels.

Originally developed in the *Gifi* system, categorical regression with optimal scaling technique (CATREG) incorporates the traditional features of multiple regression, with the exception that this method can also be applied when the predictor and outcome variables are any combination of numeric, ordinal, or nominal [10]. Following the transformation of these

nominal and ordinal variables into interval (numeric) variables using the CATREG with optimal scaling approach, multiple regression analysis is used to quantify categorical variables according to a certain scaling level.

The CARTEG model was used in this study with the aim of relating the three contributing demographic factors of age, gender, and educational level to the average employment tenure per position of jobseekers. The three converted demographic factors were taken into account as predictor variables (X_n) in this model, and the transformed variable, average job tenure per position, was designated as the outcome variable (Y). The categorical regression model is written as a linear regression equation displayed below with the use of these variables.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \quad \varepsilon$$

Where,

- β denotes the standardized regression coefficients.
- ε is a random error that has an expected value of zero and a constant variance $\sigma^{2.}$ That is, $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$ for each recorded value of the dependent variable Y.

3 FINDINGS AND DISCUSSIONS

Using the "CATREG" function, the categorical regression analysis was carried out in the IBM-SPSS software. Table 3 shows that there are no extra instances or missing values, meaning that a total of 18,664 valid examples were employed in the study.

Table 3. Case processing summary

Valid Active Cases	18664
Active Cases with Missing Values	0
Supplementary Cases	0
Total	18664
Cases Used in Analysis	18664

According to adjusted-R^2 value shown in Table 4, only 29.7% of the overall variance in the Y values around their mean can be attributed to the predictor variables included in this model. The ranges of R^2 and adjusted-R^2 are both comprehensible between 0% and 100%, and the greater these ranges are, the better the model matches the data [11]. The constrained capacity to assess the model appropriateness is one of the main drawbacks of utilizing coefficient of determination (R^2) and/or modified coefficient of determination (adjusted-R^2). Because it is challenging to forecast human behavior, this constraint arises when trying to do so. As a result, any scientific research that looks into this area typically has R^2 and/or adjusted-R^2 values below 50%.

Table 4. Model summary

Multiple R	R Square	Adjusted R Square	Apparent Prediction Error
0.545	0.297	0.297	0.703

Table 5. ANOVA test report

	Sum of Squares	df	Mean Square	F	Sig.
Regression	5542.777	5	1108.555	1576.334	0.000
Residual	13121.223	18658	0.703		
Total	18664.000	18663			

3.1 Multicollinearity diagnostics

A linear relationship between two predictor variables is known as collinearity. In the event that there is a precise linear relationship between two regressor variables, they are said to be perfectly collinear [12]. When two or more predictor variables in a multiple regression model are correlated with one another, that is referred to as multicollinearity. In order to identify predictor variables that are extremely collinear and to find solutions to the multicollinearity problem [13], researchers and academics have developed a variety of strategies. In this work, the statistical analysis of correlation coefficients and tolerance levels is the main focus.

According to [14], one of the methods researchers most frequently employ to find multicollinearity is the pairwise correlation coefficient between predictors. If two variables have a strong correlation, a regression model that includes both of them may become multicollinear. Multicollinearity may exist when the correlation coefficient is higher than 0.8, according to [15], who also claimed that this is the most common cutoff when utilizing correlation coefficients. While some utilize correlation coefficient cutoffs of 0.5 and higher [16].

Table 6. Correlation matrix for original variables

	Age	Gender	Educational Level
Age	1.00	-0.135	0.193
Gender	-0.135	1.000	0.080
Educational Level	0.193	0.080	1.000
Dimension	1	2	3
Eigenvalue	1.203	1.074	0.723

The level of tolerance for multiple regression models, according to [17], is estimated by $1\text{-}R^2$, which is employed as a measure of multicollinearity. The tolerance levels correspond to the value of $1\text{-}R^2$ when each predictor variable is regressed against each other predictor variable. The degree of multicollinearity is higher the closer the tolerance threshold goes to zero. To put it another way, a low level of tolerance indicates a significant amount of multicollinearity in multiple regression studies; the multicollinearity is present whenever the level of tolerance is below 0.40. The original relationship between the predictor variables (see Table 6), and the

transformed variables (see Table 7) are displayed. As we observe both the original and transformed predictor variables' correlation coefficients are below the 0.5 or 0.8 cutoff values. Table 8 also clearly demonstrates that all tolerance levels—both for the original predictor variables and the transformed predictor variables—are more than 0.4. Therefore, we draw the conclusion that there is no risk of our estimation module deviating from the regression assumption that there is no multicollinearity.

Table 7. Correlation matrix for transformed variables

	Age	Gender	Educational Level
Age	1.00	0.135	0.200
Gender	0.135	1.000	-0.084
Educational Level	0.200	-0.084	1.000
Dimension	1	2	3
Eigenvalue	1.209	1.076	0.715

3.2 Regression coefficients and regression equation

When all other predictor variables are maintained constant, regression coefficients—also known as slope coefficients—indicate the mean change in the value of the outcome variables that corresponds to the unit change in the predictor variable. Age, gender, and educational level are shown as the significant predictor variables of the outcome variable, which are statistically significant at the level of 0.05, according to the regression coefficients shown in Table 9. All three predictor variables have a positive relationship with the outcome variable, which means that the longer the predicted employment tenures are, the older the job candidates or employees are and the greater their educational attainment. Unlike age and educational level, gender is a nominal categorical variable and does not show the gradient effect relationship with the outcome variable. In addition, based on the data in Table 8, we can say that age is the best predictor out of the three. The estimated categorical regression model for anticipating employment tenures according to the age, gender, and educational level of job applicants or jobholders is presented below based on the regression study results:

$$\hat{Y} = 0.532*X_1 + 0.031*X_2 + 0.034*X_3$$

Table 8. Correlations and tolerance

	Correlations			Importance	Tolerance	
	Zero-Order	Partial	Part		After Transformation	Before Transformation
Age	0.543	0.523	0.515	0.973	0.937	0.940
Gender	0.100	0.037	0.031	0.011	0.971	0.970
Educational Level	0.139	0.040	0.034	0.016	0.949	0.951

Table 9. Regression coefficients

	Standardized Coefficients		df	F	Sig.
	Beta	Bootstrap (1000) Estimate of Std. Error			

Age	0.532	0.006	1	7516.720	0.000
Gender	0.031	0.006	1	25.807	0.000
Educational Level	0.034	0.007	3	25.780	0.000

4 CONCLUSIONS

Job tenure may be explained as a measure of job stability, and because of this, both the scholarly community and business management have taken an interest in it. An experimental understanding of the evolution of employment soundness can be gained by looking at the level and structure of this indicator across time. Examining occupation residency enables a determination of whether and to what extent employment stability has decreased over the past few decades, or whether concerns about dwindling work permanence are overblown. The findings of this investigation showed that there was a measurably significant association between an employee's age and their intention to leave their job. Furthermore, a company's employment tenure can help them enhance its internal people or human resource management processes, from talent recruiting and selection to training and development to succession planning, in addition to serving as a measure of their employees' loyalty. Employers should therefore make a deliberate effort to foster and sustain employee loyalty and job embedding.

This study adds to the body of scholarship on how individual-level traits affect job duration while also recognizing the importance of workplace factors such work-life balance, job designs, monetary and non-monetary perks, employment protection laws, and labor market institutions. The decision to remain or seek employment on the market at the individual level has been clearly stated in the literature. When it comes to estimating job tenure, current research takes a broad modeling approach. Models of the company's interest in work through occupancy destruction and work creation are less common [18], [19]. Attempts to merge these activity and specialty stream ideas into a comprehensive coordinating model across firms and representatives, in any case, are still in the experimental stage and have not yet resulted in codified hypothetical models. This work added to existing knowledge with the express goal of addressing an under-researched emerging country setting, and the CATREG model built in this study contributes to ongoing attempts to codify hypothetical modeling of employment duration.

REFERENCES

[1] Butler, S. S., Brennan-Ing, M., Wardamasky, S. and Ashley, A., “Determinants of longer job tenure among home care aides: What makes some stay on the job while others leave?” J. Appl. Gerontol. 33(2), 164–188 (2014).

[2] Gunsch, D., “Benefits leverage hiring and retention efforts,” Pers. J. (1992).

[3] Tippet, J. and Kluvers, R., “Employee rewards and motivation in nonprofit organisations: Case study from Australia,” Int. J. Bus. Manag. 4(3), 7–14 (2009).

[4] Hassan, B. and Olufemi, O., “Demographic variables and job performance: any link?” Acta Univ. Danub. Œcon. 10(4) (2014).

[5] Ballinger, G., Craig, E., Cross, R. and Gray, P., “A Stitch in Time Saves Nine: Leveraging Networks to Reduce the Costs of Turnover,” Calif. Manage. Rev. 53(4), 111–133 (2011).

[6] Shakil, M., “A multiple linear regression model to predict the student’s final grade in a mathematics class,” Sam Houst. State Univ. (2006).
[7] Tabachnick, B. G. and Fidell, L. S., “Analysis of covariance,” Using Multivar. Stat. 8(1), 321–374 (1996).
[8] Büyüköztürk, Ş., “Factor analysis: Basic concepts and using to development scale,” Educ. Adm. Theory Pract. 32(32), 470–483 (2002).
[9] Meulman, J. J., “Optimal scaling methods for multivariate categorical data analysis,” SPSS White Pap. Chic. (1998).
[10] Gifi, A., [Nonlinear multivariate analysis], Wiley-Blackwell (1990).
[11] Maydeu-Olivares, A. and Garcia-Forero, C., “Goodness-of-fit testing,” Int. Encycl. Educ. 7(1), 190–196 (2010).
[12] Akinwande, M. O., Dikko, H. G. and Samson, A., “Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis,” Open J. Stat. 5(07), 754 (2015).
[13] Montgomery, D. C., Peck, E. A. and Vining, G. G., [Introduction to linear regression analysis], John Wiley & Sons (2021).
[14] Vatcheva, K. P., Lee, M., McCormick, J. B. and Rahbar, M. H., “Multicollinearity in regression analyses conducted in epidemiologic studies,” Epidemiol. Sunnyvale Calif 6(2) (2016).
[15] Berry, W. D., Feldman, S. and Stanley Feldman, D., [Multiple regression in practice], Sage (1985).
[16] Donath, C., Gräßel, E., Baier, D., Pfeiffer, C., Bleich, S. and Hillemacher, T., “Predictors of binge drinking in adolescents: ultimate and distal factors-a representative study,” BMC Public Health 12(1), 1–15 (2012).
[17] Adeboye, N., Fagoyinbo, I. and Olatayo, T., “Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients,” IOSR J. Math. 10, 16–20 (2014).
[18] Mortensen, D. T. and Pissarides, C. A., “Job creation and job destruction in the theory of unemployment,” Rev. Econ. Stud. 61(3), 397–415 (1994).
[19] Caballero, R. J. and Hammour, M. L., “On the timing and efficiency of creative destruction,” Q. J. Econ. 111(3), 805–852 (1996).