Statistical analysis of South Korean senior workers' skill usage

April 15, 2018

Xinyao Fan – MSc. Statistic Student

Tom Hyeongcheol Park – MSc. Statistic Student

Nikolas Krstic – MSc. Statistic Student

Summary

This auxiliary analysis aims to explore how demographic, organizational and learning factors are associated with South Korean senior worker's skill usage, specifically numeracy and literacy. Since the response variables are averages of Likert scale items, we discretize the variables and use ordinal regression models. Many of the factors are significant predictors of skill usage, particularly education level. We discuss the limitations and potential extensions of this analysis.

Introduction

South Korea's population is "rapidly aging", with seniors composing a large proportion of the total population. Recent literature shows that South Korean senior workers continue to participate in the labor market (Keese, 2004). To further develop their workplace skills, seniors often engage in educational programs. Our research objective is to determine how demographic, organizational and learning factors are associated with South Korean senior worker's skill usage, specifically numeracy and literacy. It is critical to identify these factors, since skill usage of workers is a key area to improve both individual and organizational performance.

Exploratory Analysis

The variables within the dataset were computed using data from the open source Programme for the International Assessment of Adult Competencies (PIAAC). From August 2011 to March 2012, the organization collected survey data from 24 countries. Our focus is on employees in South Korea with ages between 50 and 65. The skill usage variables are computed as an average of Likert scale items (the questions asked in the survey), with each item having five distinct responses ranging from 1 to 5. Upon initial examination, it seems that the majority of average responses are (or close to) 1 (Figure 1). Due to this skewness in the responses, we divide the average responses into four distinct categories. We classify average responses between 1 and 2 as Category 1, classify average responses between 2 and 3 as Category 2, etc. These categories can be interpreted (in ascending order) as "low", "medium", "high" and "very high" skill usage. Thus, by discretizing the original variables, we improve interpretability and handle the inflation of 1s with our analytical approach. Table 1 contains a list of variables (and their descriptions) used in our analysis. Figure 2 displays boxplots of a few potentially important predictors. For many of the other predictors (not shown), there are no clear associations with the skill usage variables, at least visually.

Figure 1: Distributions of the skill usage variables after discretization into Categories 1-4

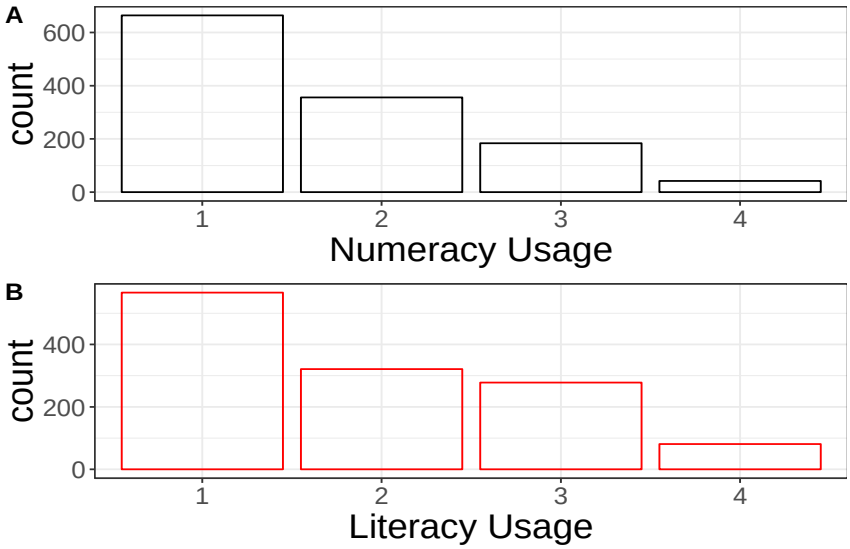
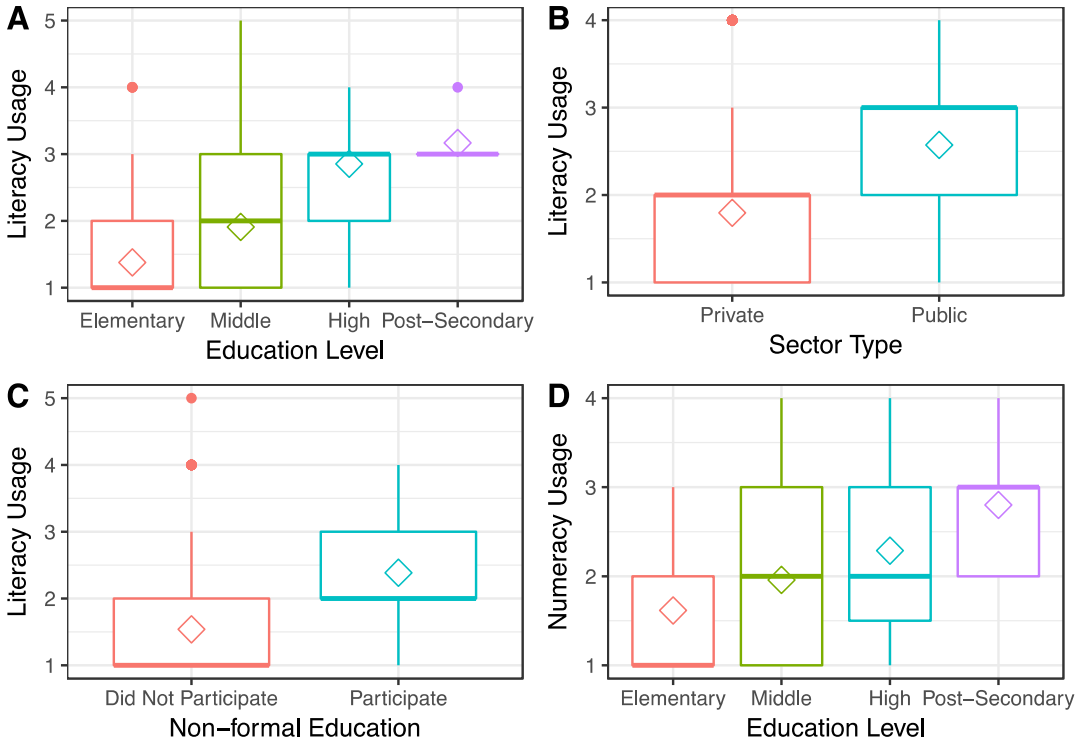


Table 1: Table of variable descriptions. Variables that do not appear here were excluded because of multicollinearity issues during modelling (high associations between predictors).

Type	Variable	Scale	Description
Demographic Factor	Age	Continuous	Age of individual. 50 to 65.
	Years of Work		Number of years at work
	Gender	Categorical	Gender. 1-male, 2-female
	Education Level		Education Level. 1-Elementary School, 2-Middle School, 3-college, 4-post-secondary
	Employment Type		Employment type. 1-fulltime, 2-part-time
Organizational Contexts	Work Flexibility	Continuous	Work flexibility (Averaging of several Likert scale items)
	Work Learning Opportunity		Participation in learning opportunities at the workplace (Averaging of several Likert scale items)
	Sector Type	Categorical	Public vs. private sector. 1-private, 2-public
	Self-employed		Whether or not individual is self-employed
Learning/ Education	Active Learning	Continuous	Active learning strategies. (Averaging of several Likert scale items)
	Non-formal Education	Categorical	Participation in non-formal education
Response	Literacy Usage	Categorical (Ordinal)	Literacy skills usage (1, 2, 3, 4)
	Numeracy Usage		Numeracy skills usage (1, 2, 3, 4)

Figure 2: Boxplots of potentially important predictor variables. The diamonds indicate the mean value while the solid line indicates the median value.



Ordinal Logistic Regression

Since the original average response variables are bounded, and are far from normally distributed, we do not support using standard linear regression due to major violations of its assumptions. To investigate the relationships between skill usage and the various factors, our approach involves modelling with an "ordinal logistic regression" (OLR). Instead of a continuous response variable, an OLR is a model for an ordered categorical response (e.g., Likert scale item). In this analysis, we use OLR for the discretized skill usage variables.

In an OLR, the model's response is in terms of log odds. Odds are essentially a ratio of probabilities. For example, if we're making the comparison of Category 1 versus Categories 2-4, then the "Odds of Category 1" is the probability of an individual being in Category 1 relative to the probability of **not** being in Category 1 (Equation 1). In terms of interpretation, a larger odds value corresponds to a higher probability of the individual being classified as Category 1. In our case, OLR models the relationships of Category 1 versus Categories 2-4, Categories 1-2 versus Categories 3-4 and Categories 1-3 versus Category 4 **simultaneously**. Once an OLR is generated, we can interpret the model's coefficients by exponentiating them to obtain odds ratios. For example, if "Age" has an odds ratio of 2, then for one unit increase in "Age" (holding every other variable constant), the "Odds of Category 1" are multiplied by 2. This means an older individual is more likely to belong to a lower skill usage category. An odds ratio

of 1 means the association between the predictor and the response is negligible. Due to the relative complexity of OLRs, we also recommend several resources for additional information on OLRs (Norušis, 2012; Fullerton, 2009).

Equation 1: Odds of Category 1. Each "p" represents the probability of an individual belonging to a category (represented by the subscript).

$$Odds\ of\ Category\ 1 = \frac{p_1}{p_2 + p_3 + p_4}$$

Final Models

To select the final models for each skill usage variable, we examine every possible combination of predictors and select the model with the lowest Akaike Information Criterion (AIC). For the numeracy model (Table 2), all predictors are included except for "Gender" and "Self-Employment". Each of the included predictors are also significant at the 5% significance level. Notably, the smallest odds ratios are from education level, suggesting that higher education is associated with greater levels of "Numeracy Usage". For the literacy model (Table 3), all predictors are included except for "Years of work". Similarly in this model, all included predictors are significant at the 5% significance level. Once again, the smallest odds ratios are from education level, but the other predictors seem to also be practically significant (their odds ratios are far from 1). When comparing the two models, a notable finding is that "Sector Type" have opposite effects on numeracy and literacy usage. Members of the public sector are associated with higher odds of literacy usage compared to private sector members, and vice versa for numeracy usage.

Table 2: Table of coefficients and corresponding odds ratios for the "Numeracy Usage" model. Odds ratios are obtained by making the coefficient the opposite sign, and then exponentiating.

Variable	Coefficient	P-value	Odds Ratio
Gender	-0.637	< 0.001	1.890
Age	-0.067	< 0.001	1.070
Education Level (Middle School)	0.840	< 0.001	0.432
Education Level (High School)	1.947	< 0.001	0.143
Education Level (Post-secondary)	1.840	0.015	0.159
Employment type (Part-time)	-0.715	< 0.001	2.045
Self-employed	-0.227	0.046	1.255
Sector type: Public	0.551	0.005	0.577
Work Flexibility	0.181	0.001	0.834
Work Learning Opportunity	0.466	< 0.001	0.627
Acitve Learning	0.648	< 0.001	0.523
Non-formal education	0.771	< 0.001	0.463

Table 3: Table of coefficients and corresponding odds ratios for the "Literacy Usage" model. Odds ratios are obtained by making the coefficient the opposite sign, and then exponentiating.

Variable	Coefficient	P-value	Odds Ratio
Age	-0.076	< 0.001	1.079
Education Level (Middle School)	0.765	< 0.001	0.465
Education Level (High School)	1.353	< 0.001	0.258
Education Level (Post-secondary)	1.590	0.046	0.204
Employment type (Part-time)	-0.405	0.042	1.500
Years of work	0.012	0.021	0.988
Sector type: Public	-0.490	0.017	1.632
Work Flexibility	0.413	< 0.001	0.662
Work Learning Opportunity	0.328	< 0.001	0.720
Acitve Learning	0.459	< 0.001	0.632
Non-formal education	0.318	0.018	0.728

Model Diagnostics

An important assumption of the OLR is the proportional odds assumption. According to this assumption, the coefficients that describe the relationships for Category 1 versus Categories 2-4 are the same as those that describe the relationships for Categories 1 and 2 versus Categories 3 and 4, and lastly the relationships for Categories 1-3 versus Category 4. We conduct likelihood ratio tests to verify the validity of this assumption in both models. For the literacy model, there is evidence to suggest that two predictors, "Active Learning" and "Work Learning Opportunity", do not satisfy the assumption. However, the p-values of these two variables are 0.027 and 0.039, which means they may only marginally violate the assumption. For the numeracy model, the p-value of "Active Learning" is much smaller than the significance level of 0.05, suggesting that the proportional odds assumption does not hold for this predictor. This implies that the effect of "Active Learning" is different when making different Category comparisons. Since OLR is also susceptible to issues of multicollinearity, elimination of highly associated predictors prior to the analysis results in successful model convergence.

Discussion and Conclusions

The literature on diagnostic or evaluation tools for OLR are relatively scarce (Abreu et al., 2009). Typically, the "goodness of fit" of an OLR is checked using Pearson or deviance tests. However, these methods are only sensible when the predictors are all categorical variables. Researchers now propose to use residual score graphs to check the proportional odds assumption and use partial residuals to check if all the predictors of the model have linear behavior. To explore these

tools, we recommend using the R package "rms" for further investigation and more complex adjustments of OLR.

Another proposed method of analyzing the data is using an inflated discrete beta regression (IDBR) (Taverne and Lambert, 2014). Since the original response variables are averages of Likert scale items, this approach is appropriate for response variables that are discrete, bounded, and have inflated values (like the 1s in our variables). Although this model seems ideal for this project, there are several concerns. These concerns include the novelty of the approach (very limited literature), extremely difficult implementation, and poor interpretability. Further details of this model can be examined in the paper by Taverne and Lambert (2014).

Our OLR models for skill usage identify several significant associations with multiple predictors. Of these predictors, "Education Level" seems to have the most prominent effect. As "Education Level" increases, the odds of an individual belonging to a low category of skill usage decreases. Similar relationships can be observed for other predictors. We also observe interesting relationships between different skill usages and "Sector Type". Given these results, further investigation into the effect of these variables on skill usage is recommended.

References

1. Keese, M. (2004, Jan). Ageing and Employment Policies in Korea – the challenge of an ageing population, *Ageing Society (OECD/OCDE)*, Jan 2004. Retrieved April 15, 2018 from http://www.eldis.org/document/A17904.

2. Norušis, M. J. (2012). *IBM SPSS statistics 19 advanced statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall. Retrieved from http://www.norusis.com/pdf/ASPC_v13.pdf

3. Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models.Sociological Methods & Research, 38(2), 306-347. doi:10.1177/0049124109346162

4. Abreu, M., Siqueira, A., & Caiaffa, W. (2009). Ordinal logistic regression in epidemiological studies. *Revista De Saude Publica*, 43(1), 183-194.

5. Taverne, C., & Lambert, P. (2014). Inflated discrete beta regression models for likert and discrete rating scale outcomes.