Factors of Numeracy and Literacy Proficiency for South Korean Senior Workers

Zhen Liu

March 7, 2018

1 Summary

The study aims to identify how demographic, organizational and learning factors are associated with South Korean senior workers' proficiency of numeracy and literacy. Stepwise AIC and adjusted R^2 are used in the statistical analysis, which indicate what factors are important to predict the numeracy/literacy proficiency test score. The result shows that numeracy and literacy proficiency test score are highly correlated. For both dependent variables, the highly important factors for them are: demographic(gender, age, education level), organizational (work flexibility, public/private sector), and learning factors (active learning strategies).

2 Introduction

The senior population in South Korea is rapidly growing. As a result, seniors continue to participate in the labour market, and often engage in educational programs to further develop their workplace skills. Identifying what is the association between the aspects of senior workers and their proficiency is a critical research issue because it will help Human Resource officers to develop training programs.

3 Data Description

The data set was obtained from the open source of the Programme for the International Assessment of Adult Competencies. These data were collected from questionnaires that measured two key cognitive skills: literacy and numeracy. Based on the study's objective, employers aged 50-65 year-old in South Korea were included. There are 1247 observations in total and two dependent variables: the average test score of numeracy proficiency(Pvnum) and the score of literacy proficiency(Pvlit). For independent variables, we have 14 independent variables that need to be test and their breif description and abbreviation are shown in Table 1:

Demographic Factor	Organizational Contexts	Learning/Education
Continuous: Age Work Experience in years(WorkY)	*Continuous*: Work Flexibility(Flex) Learning opportunity(Oppo)	*Continuous*: Active learning strategies(Active) Hours of participation in non-formal education(Hour)
Categorical: Gender Education level (EDLevel) Employment type: Full-time or part-time(Full)	*Categorical*: Private/ public sector (Priv) Manage others (Mgr) Numer of Managing employees (Mgr_c)	*Categorical*: Participation in: non-formal education (NFE) adult education (AE) job-related education (EJR) job-related adult education(AEJR) non-job-related adult education (AENJR)

Table 1: Description of independent variables

4 Method

4.1 Forward-stepwise Adjusted R^2

To find the important factors for a dependent variable, forward-stepwise selection might be a good way. The process of forward is that select best singel variable x_1 to enter model; it stays in the model while the next best x_2 is looker for; then the next best x_3 is looked for after x_1, x_2, etc. The criteria for best is various. We used the Adjusted R^2 since it helps to seek a model that has a good fit to the truth and with few parameters. It increases only if the new term improves the model more than would be expected by chance.

4.2 Forward-stepwise AIC

Using multiple approach to do model selection could provide further evidence supporting the inclusion or exclusion of certain variables. Another criteria is AIC, which increases as the goodness-of-fit is improved but decreases as the number of parameters increases. Therefore, AIC helps find a model that is a trade-off between good fit and simplicity.

4.3 ANOVA Comparison

ANOVA Comparison use R^2 as a criterion to compare if two models are the same. For example, when we compare model 1 (A+B+C) and model 2 (A+B) using ANOVA, F test is used and its corresponding p-value is measured. If the p-value for them are considerable small enough(say, < 0.05), these two models are more likely to be different so factor C could provide important information for the variable we test. On the other hand, if the p-value is not small enough, these two models do not have big difference and due to parsimony, model 2 would be a better choice.

5 Result

5.1 Data Pre-processing

We checked the missing values and odd values among these 1247 observation first. One observation has missing values for its four needed independent variables so we would exclude this observation.

There are 39 missing values in the 'Priv' variable and we have no idea which sector they are. It occupies 3% of the total observation, which is comparably small so we would exclude these 39 observations. There are 501 missing value in the 'Mgr' column and 1015 missing value in the 'Mgr_c' column, which contain both more than 40% of the total number of the observations. I would suggest that progressing our regression analysis without these two variables first.

For education level(EDLevel), there are only one observation in 'research' category and four observation in 'college' category, which might result a pretty poor estimate of the corresponding effect. It might be best to classify them as category 3(college) and regard this category as college or higher education.

We noticed that Adult education(AE) is the union of other two independent variables: Job-related adult education(AEJR) and Non-job-related adult education (AENJR). If we put them together into our models, they are probably highly correlated and the process of our variable selection might be affected. These three independent variables are all binary variables(participate or not). We tested their effects on two types of test scores and the evidence suggests that the participation of these three education all have effects on the test scores. We could either keep AE or keep AEJR and AENJR. Also, we checked that AEJR and EJR have exactly the same value. In order to avoid extremely high correlation, we would just keep one of them. Therefore, among these four independent variables(AE, AEJR, AENJR, EJR), we would keep AEJR and AENJR for following analysis.

There are 1027 observations and 13 variables are used for variable selection. Summary Statistics for categorical variables is shown in the Appendix.

5.2 Variable Selection for Literacy Score

The comparison for two possible models are shown in Table 2. Model I is the best model that forward selection picked, which has the largest adjusted R^2 and smallest AIC.

Model	Variables	Adjsted R^2	AIC
I	Gender + Age + EDLevel + Flex + Priv + Active + Hour	0.304	8402.8
II	Gender + Age + EDLevel + Flex + Priv + Active	0.303	8403.4

Table 2: Model Comparison for Literacy Score

However, the estimate of variable *Hour* in model 1 is -0.003(close to zero) and its corresponding p-value is This evidence suggests that employers with different number of hour in non-formal education probably got same scores on average in literacy proficiency test. To verify this, ANOVA comparison between Model I and Model II is processed and its corresponding p-value result is 0.12. This suggests that these two models do not have big difference; in other words, the additional variable in Model I (Hour) does not provide more information than model II. Also, their adjusted R^2 and AIC have slightly differences. Therefore, for parsimony, we would recommend Model II as the 'best' model.

5.3 Variable Selection for Numeracy Score

The comparison for three possible models are shown in Table 3. Model I is the best model under forward Adjusted R^2 criteria, which has the largest adjusted R^2 through the forward process. Model II is the best model under AIC criteria, where its AIC is the smallest throughout the forward process. The differences between these two models are NFE, AEJR, and AENJR. In model I, the p-values of Hour, NFR, AEJR, and AENJR are all larger than 0.05, which state that employees under different levels of these variables probably have same test results. Similarly, the p-value of Hour in model II is larger than 0.05.

Model	**#**	**Variables**	**Adjsted R^2**	**AIC**
I	11	Gender + Age + EDLevel + Flex + Priv + Active + Hour + NFE + AEJR + AENJR + WorkY	0.3577	8626.2
II	7	Gender + Age + EDLevel + Flex + Priv + Active + Hour	0.3564	8626.0
III	6	Gender + Age + EDLevel + Flex + Priv + Active	0.3557	8626.3

Table 3: Model Comparison for Numeracy Score

Table 4 shows the ANOVA Comparison between model III and other models in order to figure out if Hour, NFR, AEJR, and AENJR are needed. All the p-values are above 0.05, which suggests that these models do not have big differences and not provide additional effective information compared to model III. Also, their adjusted R^2 and AIC have slightly differences. Therefore, for parsimony, we would recommend Model III as the 'best' model.

Model	**Variables**	P-value
III	Gender + Age + EDLevel + Flex + Priv + Active	-
I	Gender + Age + EDLevel + Flex + Priv + Active + Hour + NFE + AEJR + AENJR + WorkY	0.16
II	Gender + Age + EDLevel + Flex + Priv + Active + Hour	0.17
IV	Gender + Age + EDLevel + Flex + Priv + Active + NFE	0.46
V	Gender + Age + EDLevel + Flex + Priv + Active + AEJR	0.75
VI	Gender + Age + EDLevel + Flex + Priv + Active + AENJR	0.31
VII	Gender + Age + EDLevel + Flex + Priv + Active + WorkY	0.40

Table 4: ANOVA Comparison for Numeracy Score

6 Conclusion and Discussion

In conclusion, gender, age, education level, work flexibility, public/private sector and active learning strategies are highly associated with both proficiency score(numeracy and literacy). The correlation between numeracy score and literacy score are 0.93, which strongly agrees with the same results they get.

From summary statistics, there are unbalanced categorical variables. Balanced classed might provide better performance and reliability of the estimate. Moreover, for the 'Priv' variables, there are 39 missing values and we excluded them, which may introduce bias in the results. The performance of our result might be better if we could use some algorithm to do data imputation, such as K-NN(using the category of its neighbor to classify it). Also, though 'Mgr' has more than 40% missing value, we still can use this variable to make further analysis. It would be interesting if we could find some association between managers and the test scores.

7 Appendix

Variables	Categories	Size	Mean for Pvlit	Mean for Pvnum
Gender	0: Male	756	255.2	248.0
	1: Female	468	241.8	231.3
Education Level	1: Middle school	501	230.0	217.0
	2: High school	440	254.8	246.6
	3: College or higher	266	280.4	280.3
Employment Type	1: Full-time	1027	251.8	243.5
	2: Part-time	180	240.7	231.7
Private/Public Sector	1: Private	1062	247.5	238.4
	2: Public	145	269.3	266.6
Non-formal Education	0: Participation	694	242.1	231.1
	1: No	513	261.0	256.1
Job-related Adult Ed	0: Participation	801	244.9	234.9
	1: No	406	260.5	255.3
Non-job-related Adult Ed	0: Participation	1097	234.9	239.9
	1: No	110	255.3	260.0

Table 5: Summary Statistics: Categorical Variables

From Figure 2, two Q-Q plots both seem on a straight line so it is fair to say that the two responses(pvnumM, pvlitM) of our sample are normal distributed. The assumption of linear regression is statisfied.

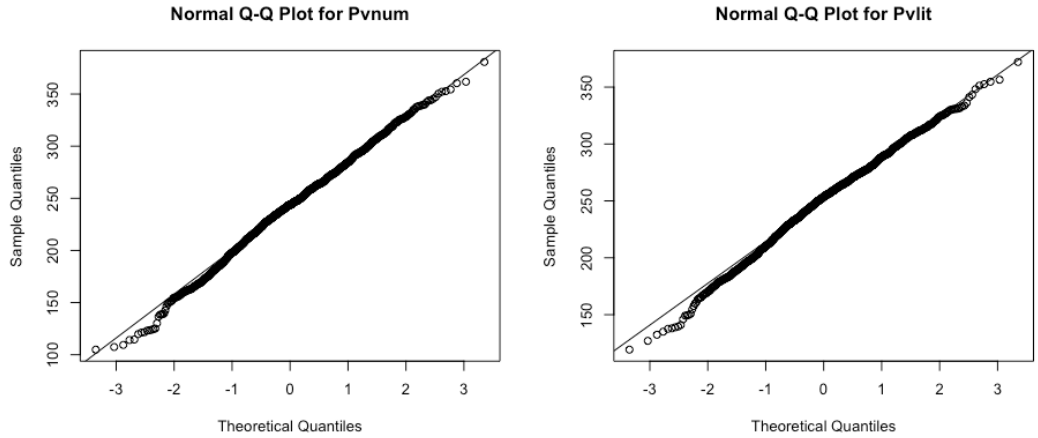

Figure 1: Q-Q Plots for dependent variables