An Analysis of The Association Between Demographic, Organizational and Learning Factors And The Literacy And Numeracy Scores Of The South Korean Senior Workers

Jingyi Huang

2018/3/5

1 Summary

In this report, we analyze how demographic, organizational and learning factors are associated with South Korean senior workers' skill usage and proficiency scores. After conducting statistical analysis, we conclude that education level, active learning strategies, age, gender, public/private sector and work flexibility are statistically significant variables that associate with both literacy and numeracy scores.

2 Introduction

As the senior population in South Korea is rapidly growing, seniors often engage in educational programs to further develop their workplace skills. Our study uses the data from the open source of the Programme for the International Assessment of Adult Competencies with 1247 observations and 25 variables. The objective of this study is to identify how demographic, organizational and learning variables are associated with South Korean senior workers' numeracy test scores and literacy test scores.

Both demographic and organizational variables are nature of the senior workers. They are usually hard to change. Hence, our client is more eager to investigate the association between learning variables and the scores that the senior worker receive. The learning process is nurtured, and we anticipate that the more learning hours the old workers experienced, the better they perform in their work. Identifying what is the association between the aspects of senior workers and their skill use is a critical research issue because it will help Human Resource officers to develop training programs.

The report presents an analysis of the association between different variables and the two types of scores mentioned in the beginning. It turns out that one of the learning variable: act_lrn is statistically significant with a positive estimate. Hence, our result agrees with our anticipation. Other variables including education level, age, gender, public/private sector, work flexibility are significantly associated with the numeracy score and literacy score.

3 Data Description

Our study uses the data from the open source of the Programme for the International Assessment of Adult Competencies with 1247 observations and 25 variables.

3.1 Responses (Dependent variables)

There are four response variables in our dataset denoted by num_use, liter_use, pvnumM and pvlitM.

Dependent Variable Name	Dependent Variable Description
num_use	Skill use for work: the frequency the worker uses certain numeracy skills (numeric)
lit_use	Skill use for work: the frequency the worker uses certain literacy skills (numeric)
pvnumM	Proficiency test scores: the numeracy test score (numeric)
pvlitM	Proficiency test scores: the literacy test score (numeric)

Table 1: This table shows the dependent variables in the data.

According to Table 1, there are four dependent variables: Skill usage for work (numeracy or literacy) and Proficiency test score (numeracy or literacy) denoted by num_use, liter_use, pvnumM and pvlitM. The skill usage for work states the frequency that numeracy/literacy skills are applied. It can be seen from Table 2 that there are 14 meaningful independent variables in the dataset after discussing with our client.

Below are four Normal Q-Q plots for the four responses.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.667   1.955   2.500   5.000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.333   2.083   2.267   3.146   5.000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   104.9   213.9   243.7   242.0   270.7   380.7

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   119.3   226.1   253.7   250.4   275.6   372.1
```

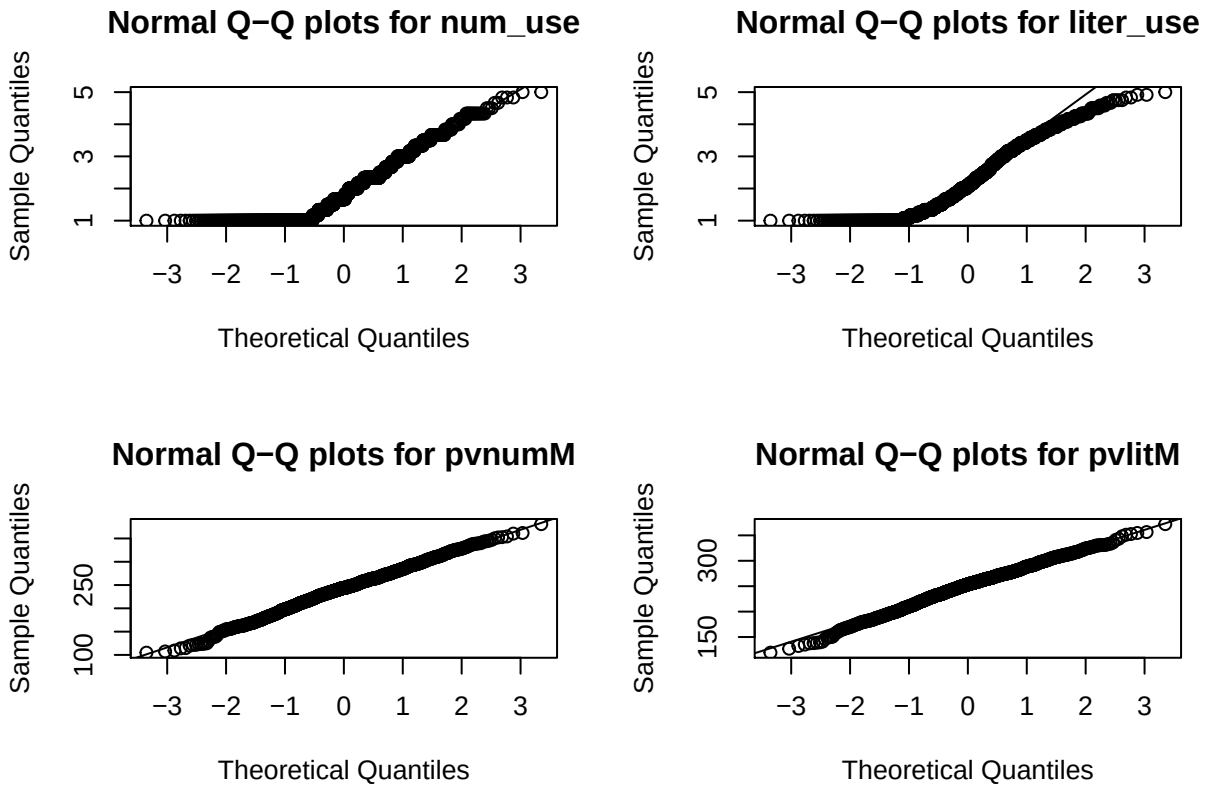

Figure 1: Normal Q-Q plots of the four responses.

There are many repeated values in num_use and liter_use that might cause the non-normality of the data. The top left and top right plots in Figure 1, indeed, do not follow straight lines. Hence, num_use and liter_use violate the normality assumption, and STAT 550 will be responsible for these two responses. The rest of the two plots seem to follow straight lines so the two responses: pvnumM and pvlitM are normally distributed. Since the assumption of linear regression is satisfied, we, STAT 450, focus on these two responses.

3.2 Independent Variables

As shown in Table 2, there are 14 independent variables. We find that FNFE12JR and FNFAET12JR are exactly the same, so we can delete one of them: FNFE12JR. FNFAET12 is the sum of FNFAET12JR and FNFAET12NJR. Using t-tests, FNFAET12, FNFAET12JR and FNFAET12NJR are all statistically significant at 5% significance level. The normality assumption is valid here. Therefore, we decide to keep FNFAET12JR and FNFAET12NJR.

3.3 Missing Values

One of the challenges is to clean the data with missing values and wrong input.The max of num_use is 5, so 7 may be a wrong input. Also, there are other four missing values in this observation. Therefore, we use the dataset that excludes these data.There are 501 missing values in the Mgr column and around 1051 missing values in Mgr_c column, which is more than 40% of the total number of the observations. Hence, we decide to delete this whole variable. Besides, there are 39 missing values in the 'pub_priv', which is around 30% of the total number of the observations. We omit the missing values in the data as usual. After deleting the observations with unknown sector, 1062 of the observations are from the private sector while 145 of them are from the public sector.

Demographic Variable Name	Demographic Variable Description	Organizational Variable Name	Organizational Variable Description	Learning Variable Name	Learning Variable Description
AGE_R	Age for each individual senior worker (numeric)	work_flexM	Work Flexibility (numeric)	act_lrn	Active learning strategies number of hour of participation (numeric)
Years_wk	Work Experience in years (numeric)	work_lrnM	Learning opportunity (numeric)	NFEHRS	Number of hour of participation in non-formal education (numeric)
GENDER_R	Gender: 1-male, 2-female (factor)			NFE12	Participation in non-formal education: 0 for no, 1 for yes (factor)
ED_Level	Education level (factor): 1-middle, 2-high, 3-college, 4-graduate			FNFAET12JR	Participation in formal or non-formal adult education program (job-related): 0 for no, 1 for yes (factor)
EMPloyed	Employment type: 1-full time, 2-part-time (factor)			FNFAET12NJR	Participation in formal or non-formal adult education program (non job-related): 0 for no, 1 for yes (factor)
				FNFAET12	Participation in formal or non-formal adult education program: 0 for no, 1 for yes (factor)
				FNFE12JR	Participation in formal or non-formal education(job-related): 0 for no, 1 for yes (factor)

Table 2: This table shows the independent variables in the data.

3.4 Two datasets for two responses

After cleaning the data, we subset our data to 2 datasets: one for the numeracy response, the other for the literacy response, each with 1207 observations and 14 variables. The scope of the two datasets is both public and private sectors.

3.5 Correlations

We check the correlations between categorical independent variables using Chi-square test and Fisher test and find that only FNFAET12JR and NFE12 have very strong correlation (r=0.8).

```
FALSE Loading required package: corrplot
```

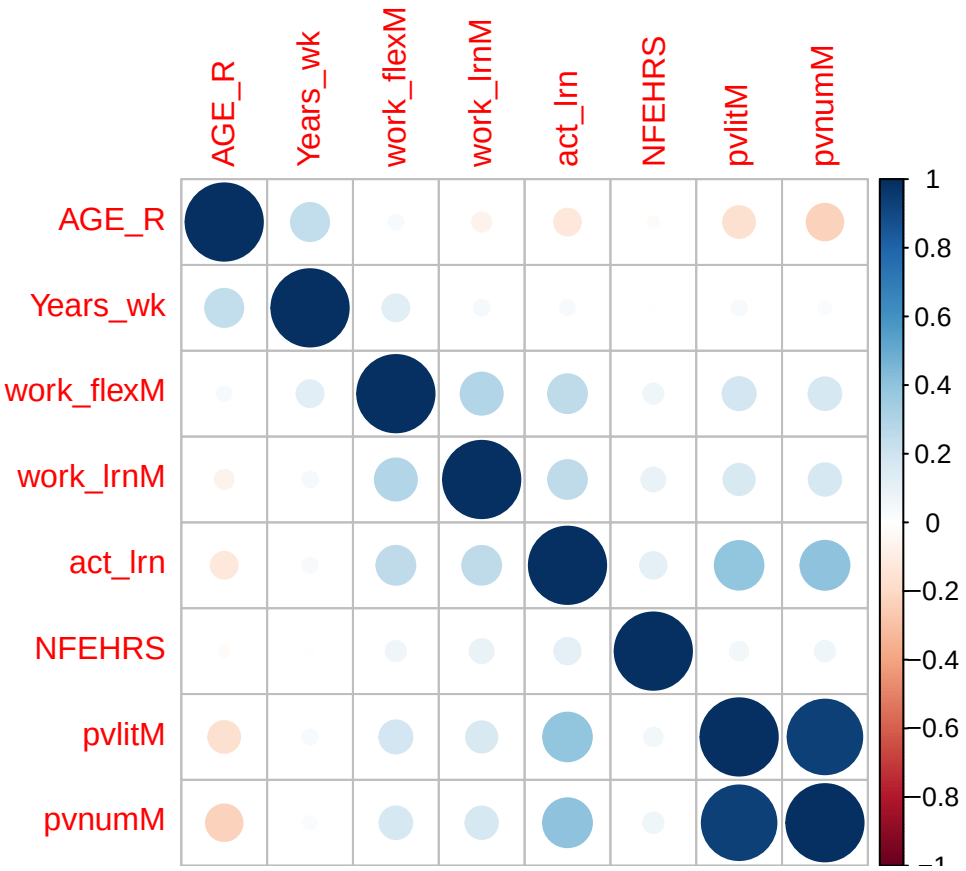

Figure 2: Correlations between numeric variables

The correlations between numeric variables are checked as well. Figure 3 shows that pvlitM and pvnumM are highly correlated (r=0.9).

3.6 Boxplots

We use boxplots to visualize our data. Since there are too many variables in our dataset and putting all of them in the report be overwhelming for our client, I choose act_lrn as an example to interpret the boxplots.

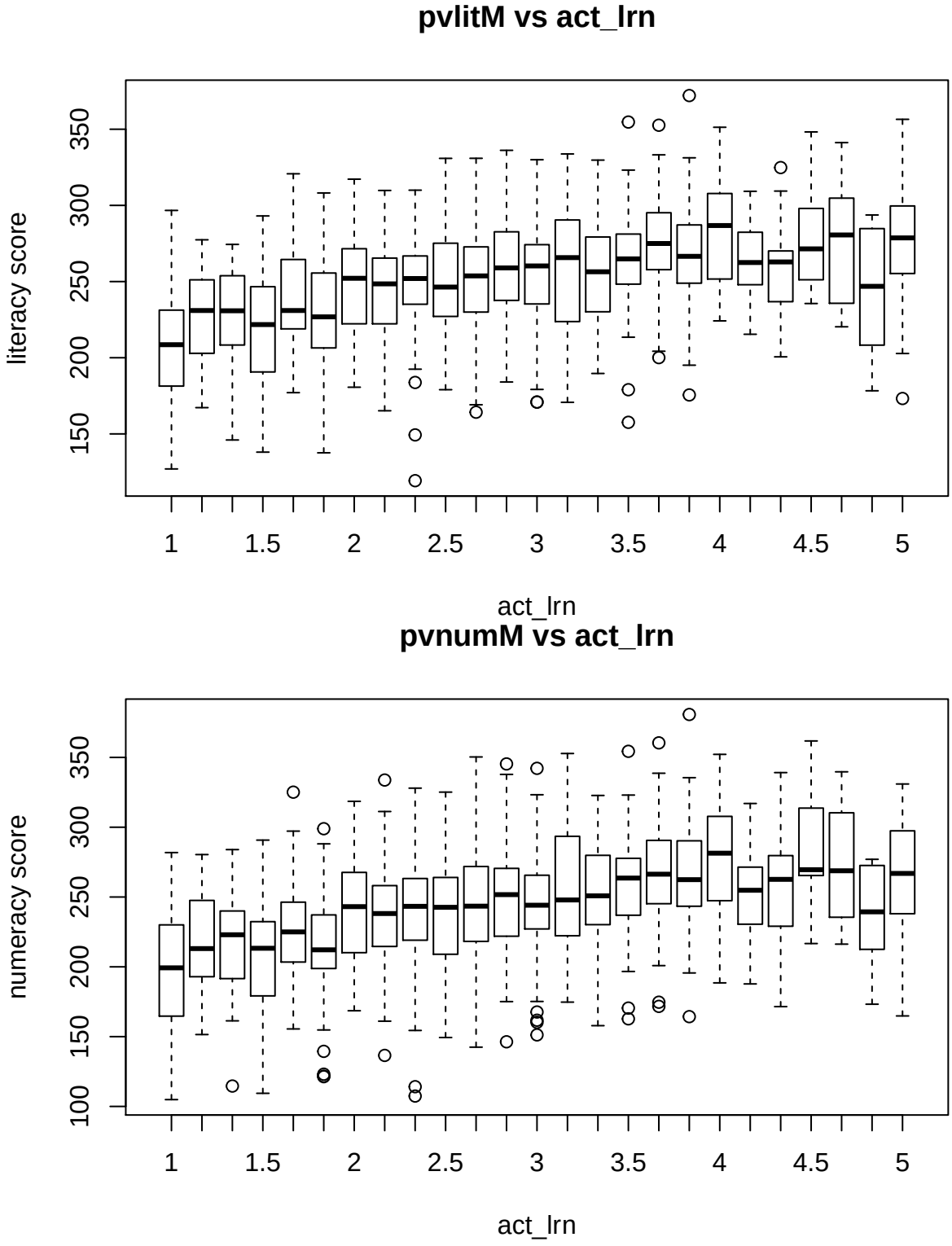

Figure 3: Boxplots for active learning strategies number of hour of participation with 2 responses

In Figure 3, there is an increasing trend between act_lrn and the 2 scores with some mild fluctuations because the median of scores rises as act_lrn increases. Hence, we expect that there is a positive association between act_lrn and the 2 scores, which is further confirmed later. The sizes of each box vary indicating that the

spread of the literacy and numeracy scores for each act_lrn are not the same. Larger box size implys larger variance of the scores.

4 Methods

Before calculating the estimates and p values of the variables, we want to select the best model first. Model selection is a process where we test to determine whether all the variables are important to keep in the model. The "full model" is defined as the model that uses all the variables after cleaning our data. To determine if these variables are significant, we compare the full model to a model where some variables are excluded. A method to compare these models is to apply Forward Adjusted R^2, Stepwise AIC and ANOVA between the models. We start with analysing literacy score.

4.1 Model Selection using Forward Adjusted R^2

Our client is most familiar with R^2. However, R^2 increases every time you add an independent variable to the model. Adjusted R^2 is better than R^2 because it penalized the number of variables adding in the model, which makes our analysis more precise. The larger the Adjusted R^2, the better the model will be. Among the 15 steps, the 7th step maximizes the Adjusted R^2 and the corresponding model is the best model in our case. As a result, GENDER_R, AGE_R, ED_Level(3), pub_priv, work_flexM, act_lrn and NFEHRS are selected by using this method (adjusted R^2=0.3).

4.2 Model Selection using Stepwise AIC

We use AIC as a criterion to select the best linear regression model because we want to assess the prediction performance of the model. AIC helps to seek a model that has a good fit to the truth but few parameters. The smaller the AIC, the better the model will be. We apply the AIC to the whole dataset including both private and public sectors for literacy score and numeracy score respectively. The direction of stepwise AIC does not affect the variables selected for the best model in our case, so we just use the forward direction. Forward means adding the variable one by one from the starting model with no variable while backward means deleting the variable one by one from a full model with all the variables. Stepwise AIC shows that ED_Level, act_lrn, Age, Gender, work_flexM, NFEHRS and pub_priv are picked by AIC to form the best model.

4.3 Model Selection using ANOVA

Then, we use another method of model selection: ANOVA to compare our results with those in stepwise AIC. Since our client is unsure about whether the study is focused on only the private sector or both public and private sectors. We first check the pub_priv variable to see if it is needed to include in our model. The p-value of pub_priv is less than 0.05. Hence, it is important to keep this significant variable in our model. We also check other variables picked by stepwise AIC and Adjusted R^2 and find that pub_priv,AGE_R,ED_Level,work_flexM, Gender_R and act_lrn are statistically significant at 5% significance level ($p<0.05$). This means that model with and without these variables will be so different that we should not discard this variable. Hence, we should keep the variables with small p-values.

Similarly, we follow the steps as above and get the similar result for numeracy score response. The best model still consists of ED_Level, act_lrn, AGE_R, work_flexM, GENDER_R and pub_priv. These variables are also significant at 5% significance level when using the ANOVA comparsion.

5 Results

After using stepwise AIC for the literacy score response, we find that the best model for the private and public sectors dataset contains 6 variables: education level, active learning strategies, age, gender, public/private sector and work flexibility. Though stepwise AIC selects NFEHRS as one of the important variables, ANOVA indicates that it is not statistically significant at 5% significance level (p=0.1). Other than that, all the variables picked by stepwise AIC are the same as the ones from using ANOVA. Hence, the results from using different model selection methods select roughly the same variables.

Coefficient	Estimate for literacy	P-value for literacy	Estimate for numeracy	P-value for numeracy
Intercept	238.503	<0.05	259.563	<0.05
ED_Level2	19.468	<0.05	22.542	<0.05
ED_Level3	36.109	<0.05	46.570	<0.05
ED_Level4	53.779	<0.05	57.460	<0.05
act_lrn	8.386	<0.05	8.653	<0.05
AGE_R	-0.564	<0.05	-1.158	<0.05
work_flexM	1.932	<0.05	1.907	<0.05
GENDER_R2	-4.062	<0.05	-5.635	<0.05
pub_priv2	6.805	<0.05	9.322	<0.05

Table 3: This table shows the estimated values for the coefficients and their associated p-value.

It is clearly shown in Table 3 that all the variables selected are significant at 5% significance level ($p<0.05$). ED_Level has positive coefficients which means that the higher the education level the senior workers have, the higher the score they receive. Similarly, act_lrn, work_flexM and pub_priv can also improve the score. Moreover, since our client is most interested in the Learning variables, my interpretation focuses more on the only significant variable left in the Learning variables: act_lrn. According to Table 3, we observe that the estimated coefficient for act_lrn is 8.386 for literacy score response. Recall that act_lrn denotes the active learning strategies number of hour of participation. Therefore, if we hold all other variables constant, and increase act_lrn by 1 hour, the literacy score increases by 8.386. A similar conclusion can be drawn from the numeracy score response as well.

6 Conclusions and discussions

We conclude that education level, active learning strategies, age, gender, public/private sector, work flexibility and number of hours of participation in non-formal education are significant factors that associate with the usage and proficiency scores. The more number of hours that the senior workers participate in active learning strategies, the higher their literacy and numeracy scores are.

We only test 2 of the 4 responses that the client gave us because only 2 of them satisfy the linear regression assumptions. These 2 responses are highly correlated which support our analysis that the significant variables are the same for these 2 different responses. In the future, we will combine our results with STAT 550 who are responsible for the other 2 responses.