Senior worker report

Linda Liu

2/27/2018

Summary

It is good to have Summary takes 1/3 to 1/2 of a page, more focusing on data analysis, result and finding part. How about add some sentences about the parts?

The purpose of the analysis is to identify how demographic, organizational and learning factors are associated with numeracy and literacy scores of South Korean senior workers. After applying model selection methods, result shows that gender, education level, work flexibilty, active learning, public or private sectors are important variables that are assoiciated with both numeracy score and literacy scores of senior workers in South Korean.

Specify methodology: What kinds of model is used?

Korea

Summary should stands on its own. How about specifying the meaning of the variables so that people who don't know what active learning can understand what you mean? :)

Introduction

Many senior works in South Korea still participate in the labour market due to the reason ~~pid aging.~~ Knowing which factors are assoicated with senior workers' skills at the first place will help improve the performance of senior worker. T the objective of this study is to identify how demographic, organizational and learning factors are associated with numeracy and literacy skills of South Korean senior workers, ~~which is also the interests of our client.~~ I guess we could erase this sentence

In the next part of the report, I include the description of the data, methods used to test the hypothesis and the corresponding results.

Data Description

I think the data set is better expression, as 'these data' may lead to a confusion of having several dataset.

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 on the ~~private sector~~ in South Korea were included. The dataset contains 1247 data entries.

measure

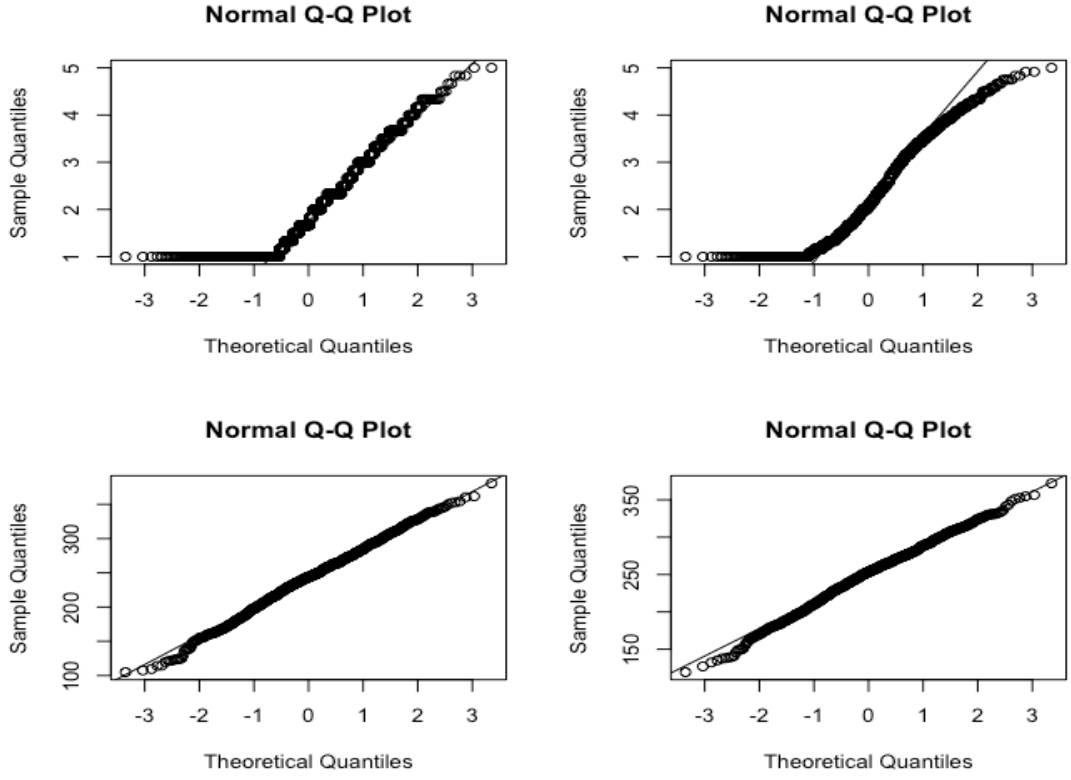
private + public sector =1247 data.

The raw dataset has 4 response variables ~~=~~ num_use, lit_use, pvnumM and pvlitM. Num_use and lit_use are the frequency of the ~~utilization~~ of certain Numeracy and literacy skills. They are calculated by taking the average of some sub-items which are mearued by likert scale: 1,2,3,4,5. The top two plots in Figure-1 show that num_use and lit_use do not follow normal distributions. PvnumM and pvlitM are proficiency test scores for numeracy and literacy. The bottom two plots in Figure-1 show that pvnumM and pvlitM follow normal distributions. Since handling response variables that do not follow normal distribution

usage; recommended to use plain English :)

needs further research, we will only focus on the analysis of pvnumM and pvlitM in this report.

I think this data description part above is well written. But to add a comment, it can be nicer if you specify which QQ plot is for which variables by adding labels.



If you used Microsoft Word, you may add some comments on the name of the Figure 1. eg. Figure 1: QQ plots for 4 response variables.

Figure-1

There are three different categories of explanatory varaibles.

1. Demographic factors:
   - Age[AGE_R]
   - Gender[Gender_R]: 1-male, 2-female
   - Education[Education Level]: 1-~~middle~~, ~~2-high~~, 3-college, 4-graduate

     Middle school and high school
   - Employment type[full_part]: 1-fulltime, 2-part-time
   - Work experiences (years) [Year_wl]: continuous
   - Public vs. private sector [pub_piv]: 1-private, 2-public
2. Organizational contexts:
   - Work flexibility[work_flexM] - average of four likert scores: sequence of tasks, how to do the work, speed of work and working hours
   - Learning opportunity[work_lrn] - average of three likert scores: learning from peers/supervisors, learning by doing, keeping up to date
   - Managing others[Mgr]: 1-yes, 2-no
   - Managing how many[Mgr_c]: 1:0'-5, 2:6-10, 3:11-24, 4:25-99, 5:more than 100

3. Learnging/Eduaction:
   - Active learning strategies[Act_lrn] - average of sub-items:relate new ideas, learning new things, attribute something new, bottoming of difficult things, fit different ideas together, look additional info(using likert scale-1,2,3,4,5)
   - Participation in non-formal education [NFE12]: 0-not; 1-yes
   - Participation in formal or non-formal adult education program [FNFAET12]:0-no; 1-yes
   - Participation in formal or non-formal education(job-related) [FNFET12JR]:0-no; 1-yes
   - Participation in formal or non-formal adult education program (job-related)[FNFAET12JR]: 0-no; 1-yes
   - Participation in formal or non-formal adult education program (non job-related)[FNFAET12NJR]:0-no; 1-yes
   - Number of hours of participation in non-formal education[NFEHRS]: continuous

One of the issures ~~One issue~~ about the dataset is that there are 501 missing ~~value~~ values in the 'Mgr' column and 1015 missing ~~value~~ values in the 'Mgr_c' colum. It is reasonable to ignore Mgr and Mgr_c because almost half of the values are missing. Also, it was observed that column FNFE12JR is identical to FNFAET12JR. For the sake of simplicity, we removed column FNFE12JR.

Figure-2 shows the correlation among all continuous variables. There is a strong correlation betweem pvlitM and pvnumM. Other correlations are either moderate or low.

Figure-3 shows the correlation among all categorical variables. There is a strong correlation between NFE12 and FNFAET12JR. Other correlations are either moderate or low.

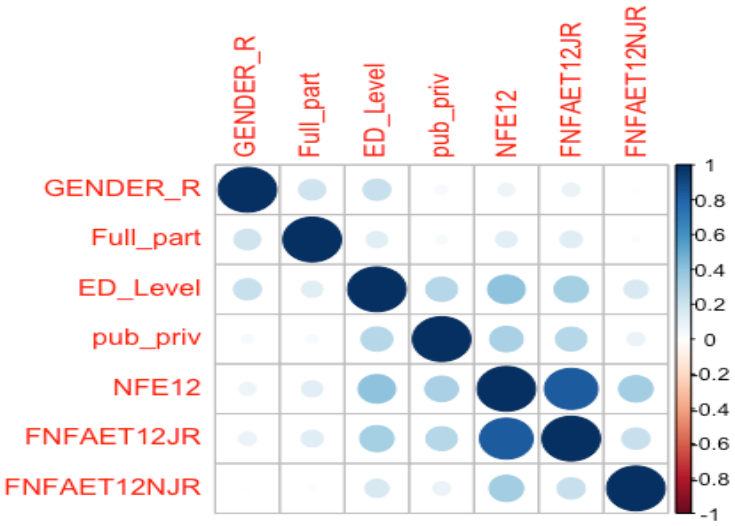


Figure-2

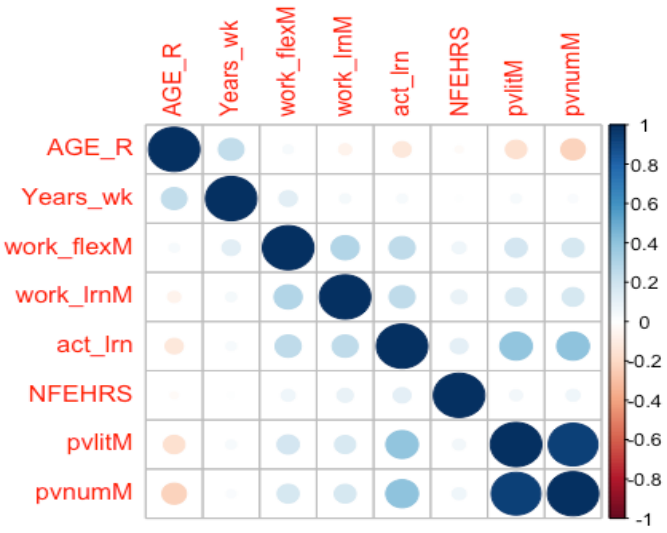


Figure-3

Methods

Methods being used to do variable selection are stepwise AIC and stepwise adjusted R-square. Stepwise means to select variables sequentially one by one. Two types of stepwise methods are used here: forward selection and backward selection. In terms of forward selection, it starts with null model (a model with no variables), then among all variables, it finds the one that is the best to improve the value of criterion. Criterion is AIC if using stepwise AIC, and criterion is adjusted R- square if using stepwise adjusted R- square . It keeps adding new variables until no more variables could be added to improve the criterion. Backward stepwise uses the same technique, but it starts from the full model (a model with all variables) and take one variable out of the model until no variables can be removed.

To verify the result, nested anova test is being used. Nested criterion allows us to compare one model with a few variables (for example A+B+C) to another model with the same variables and an additional variable (for example A+B+C+D).

Results

1. Results for numeracy score

Stepwise adjusted R-square picks GENDER_R, AGE_R, ED_Level, pub_priv, work_flexM, act_lrn, NFEHRS, NFE121, FNFAET12JR1 and FNFAET12NJR1. AIC picks the same variables except for NFE121, FNFAET12JR1 and FNFAET12NJR1. Nested avova was applied afterwards to test whether these three varaibles should be included in the model or not.

Result shows the p-values for all the three varaibles are greater than 0.05. Thus, none of these three variables is significant.

Then, we run model diagnostic by fitting a linear regression model with the selected variable: GENDER_R, AGE_R, ED_Level, pub_priv, work_flexM, act_lrn, NFEHRS.

The adjusted R-squared of the final model is 0.356, which means the picked variables could explain 35.6% of the variances in numeracy score. Table-1 summarizes the coefficient of all selected variables. The p value of each coefficient indicates the significance of each variable. The samller the p-value is, the more significant the variable is. Thus, education level and act_lrn are the top variables. NFEHRS has a p-value greater than 0.05. Thus, it is an optional variable, meaning that whether included it or not is up to the customer.

Overall, GENDER_R, AGE_R, ED_Level, pub_priv, work_flexM, act_lrn, NFEHRS (optional)are variables that are associated with numeracy score.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 259.04415   14.65463   17.68  < 2e-16 ***
ED_Level2    22.60912    2.44159    9.26  < 2e-16 ***
ED_Level3    47.16264    3.21517   14.67  < 2e-16 ***
ED_Level4    57.56368   14.79274    3.89  0.00011 ***
act_lrn       8.73426    1.20946    7.22  9.1e-13 ***
AGE_R        -1.15477    0.24545   -4.70  2.8e-06 ***
work_flexM    1.94509    0.81360    2.39  0.01697 *
GENDER_R2    -5.47547    2.18280   -2.51  0.01226 *
pub_priv2     9.67951    3.32811    2.91  0.00370 **
NFEHRS       -0.00328    0.00221   -1.49  0.13733
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table-1

2. Results for literacy score

Stepwise adjusted R-square gives the same result for both backward, forward direction. It picks gender, age, ED_level, pub_priv, work_flexM, act_lrn, NFEHRS. Stepwise AIC also picks GENDER_R, AGE_R, ED_Level, pub_priv, work_flexM, act_lrn, NFEHRS. Forward and backward direction gives the same result too.

Since stepwise AIC and stepwise adjusted R-square lead to the same result, so we conclude that GENDER_R, AGE_R, ED_Level, pub_priv, work_flexM, act_lrn, NFEHRS are variables that are associated with literacy score.

Same as literacy score, we run model diagnostics by fitting a linear regression model with the selected variable. The adjusted R-squared of the final model is 0.304, which means the picked variables could explain 30.4% of the variances in literacy score. NFEHRS again has a p-value greater than 0.05. Thus, it becomes an optional variable for literacy score too.

Conclusions

Education level, gender, age, public vs private sector, work flexibility, active learning strategies are associated with both numeracy and literacy score.