Senior worker report

Linda Liu

2/27/2018

Summary

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. By applying model selection methods, result shows that gender, education level, work flexibilty, active learning, public sector or private sector are important varibles that are assoiciated with both numeracy score and literacy scores of senior workers in South Korean. There is one additional variable that is associated with numeracy score, that is age.

Introduction: write a brief description of the problem, the objective(s) of the study, the question(s) of the client, and a brief overview of your analysis.

Introduction

Many senior works in South Korea still participate in the labour market due to the reason of rapid aging. Knowing which factors are assoicated with senior workers' skills at the first place will help improve the performance of senior worker.Thus, 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.

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

Data Description: state any relevant details about the collection of the data or measurement of the variables, and use clear tables and figures for summarizing the data.

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 on the private sector in South Korea were included.

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

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

    - 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

By obeserving the summary statistics, an unusal value was found. The max of num_use is 5 but one of the data entry has num_use of value 7, which indicate that the data entry may not have the accurate values. Also, there are other four missing values in this observation. Thus, this data entry is excluded from the whole dataset. Another issue about the dataset is that there are 501 missing value in the ‘Mgr’ column and 1015 missing value in the ‘Mgr_c’ colum. It is reasonable to ignore Mgr and Mgr_c because almost half of the values are missing.

Figure 2 below 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 below shows the correlation among all categorical variables.There is a strong correlation betweem NFE12 and FNFAET12JR Other correlations are either moderate or low.

Methods: describe the statistical methods you used in the analysis. Avoid formulas and include examples if needed. *Remember your audience here!* Do *not* discuss overly technical details that will not have substantive value to the client (remember that your client is not a statistician).

Methods

Methods being used to do variable selection are stepwise AIC and stepwise R^2. Stepwise means to select variables sequentially.Consider,if there are p covariates, then there are 2^p possible models to consider. If p is large, then it is computationally infeasible. Stepwise method helps to reduce the complexity. There are three types of stepwise method - forward, backward and both. In terms of forward selection, it starts with null model (a model with no variables), then among all p variables, it finds the one that is the best to improve the value of crietion. which in this case AIC is the crietion. Let’s say x3 is first added to the model, then it tries to find the second variable that improve the AIC. It keeps doing this until no more variables could be added to improve the AIC. Backward stepwise AIC is starting from the full model (a model which includes all varibles) and take the one that after removing it AIC improves the most.

To verify the result, nested anova test is being used.

Results: describe and explain the results of your analysis to your client! For example, if you are fitting a regression, interpret the most relevant coefficients in the context of the project. Make nice and clear tables and figures to support your points. Add clear captions and legends to them. Number your Figs and Tables to refer to them in the text.

## Results

```
# library(leaps)
# options(digits=5)
# adjr_lit<-regsubsets(pvlitM~.,data=data_lit,nbest=1,method = c("forward"),nvmax=15)
# summ.adjr_lit<-summary(adjr_lit)
# which.max(summ.adjr_lit$adjr2)
# summ.adjr_lit # GENDER_R, AGE_R, ED_Level(3), pub_priv, work_flexM, act_lrn, NFEHRS
```

```
# #literacy: use data_lit
# table(newdata$pub_priv)
# library(MASS)
# full.lit<-lm(pvlitM~.,data=data_lit)
# null.lit<-lm(pvlitM~1,data=data_lit)
# (step(null.lit,scope = list(lower=null.lit,upper=full.lit),direction="forward", trace = T))$coefficie
# #GENDER_R + AGE_R + ED_Level + pub_priv + work_flexM + act_lrn + NFEHRS, same as above
```

1. Result for literacy score Stepwise adjusted R^2 give the same result for both backward, forward direction. It picks 9 variables, including GENDER_R, AGE_R, ED_Level2, ED_Level3, ED_Level4, pub_priv, work_flexM, act_lrn, NFEHRS. ED_level is a categorical variable which has four levels - ED_Level1, ED_Level2, ED_Level3, ED_Level4. ED_Level1 is not selected because it is the baseline. ED_Level2, ED_Level3, ED_Level4 are selected, it means that ED_level is a significant variable. Thus, overall stepwise adjusted R^2 picks GENDER_R, AGE_R, ED_Level(3), pub_priv, work_flexM, act_lrn, NFEHRS

Stepwise AIC also picks GENDER_R, AGE_R, ED_Level(3), pub_priv, work_flexM, act_lrn, NFEHRS. Forward and backward direction gives the same result too.

Since stepwose AIC and stepwise adjusted R^2 lead to the same result, so we conclude that GENDER_R, AGE_R, ED_Level(3), pub_priv, work_flexM, act_lrn, NFEHRS are variables that are assciated with literacy score.

2. Result for numeracy score Stepwise adjusted R^2 picks GENDER_R, AGE_R, ED_Level, pub_priv, work_flexM, act_lrn,NFEHRS, NFE121, FNFAET12JR1 and FNFAET12NJR1. AIC picks the same varaibles except for NFE121,FNFAET12JR1 and FNFAET12NJR1. Nested avova was applied here to test whether these three varaibles should be included in the model or not. # {r} # anova(lm(formula = pvnumM ~ ED_Level + act_lrn + AGE_R+ work_flexM + GENDER_R +pub_priv, data = data_num),lm(formula = pvnumM ~ ED_Level + act_lrn + AGE_R + work_flexM + GENDER_R + pub_priv +FNFAET12NJR, data = data_num))$‘Pr(>F)‘[2] #FNFAET12NJR: 0.29617 # # anova(lm(formula = pvnumM ~ ED_Level + act_lrn + AGE_R+ work_flexM + GENDER_R +pub_priv, data = data_num),lm(formula = pvnumM ~ ED_Level + act_lrn + AGE_R + work_flexM + GENDER_R + pub_priv +NFE12, data = data_num))$‘Pr(>F)‘[2] #NFE12: 0.47319 # # anova(lm(formula = pvnumM ~ ED_Level + act_lrn + AGE_R+ work_flexM + GENDER_R +pub_priv, data = data_num),lm(formula = pvnumM ~ ED_Level + act_lrn + AGE_R + work_flexM + GENDER_R + pub_priv +FNFAET12JR, data = data_num))$‘Pr(>F)‘[2] #FNFAET12JR: 0.78678 # Result shows the p-values for all the three varaibles are greater than 0.5. Thus,none of these three varaibles is signficant. # {r} # summary(lm(formula = pvnumM ~ ED_Level + act_lrn + AGE_R+ work_flexM + #   GENDER_R +pub_priv+NFEHRS, data = data_num)) # The adjusted R-squared of the final model is 0.356, which means the picked variables could expain 35.6% of the variances in numeracy score. 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 edcation level and act_lrn are the top varaibles. NFEHRS has a p-value greater than 0.05. Thus it is a optional variable, meaning that either 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 assciated with numeracy score.

Conclusions

GENDER_R, AGE_R, ED_Level(3), pub_priv, work_flexM, act_lrn, NFEHRS are variables that are assciated with literacy score.