Predicting Where a Teacher is Likely to Teach by Sex and State

Fadlan Arif

21/12/2020

Abstract

In Malaysia, there is a split in the secondary school system. There is the standard academic school and then there is the vocational school. Each school has their own audience and people in Malaysia have a certain perspective about whose who go to a vocational school. In this paper we investigate this perception affects the likelihood of teachers joining these schools. This information could be vital in assessing how we incentivize people to work for different schools.

Within this paper we use R Core Team (2020), Allaire et al. (2020), Ushey et al. (2020), Wickham et al. (2019), Wickham and Bryan (2019), Xie (2015)

Introduction

We should first define what is a vocational school compared to and academic school before we are able to better examine the reason for our investigation. An academic school follows the same structure of primary school in which students must learn a wide variety of subjects that are more theoretical. This form of teaching is geared towards preparing the students for jobs in the tertiary sector. According to Wikipedia: 'A vocational school is a type of educational institution, which, depending on the country, may refer to either secondary or post-secondary education designed to provide vocational education, or technical skills required to complete the tasks of a particular and specific job' (*Vocational School* 2020). For Malaysia, it is offered as a secondary school option and those who chose this route have to defer from their academic school. This type of schooling prepares students for work in the primary or secondary sector. We have chosen to look into the distribution of teachers within Malaysia, as the country itself has an interesting economic status (*Malaysia: Distribution of Employment by Economic Sector from 2010 to 2020* 2020). Malaysia is considered a Developing to Developed Nation. So, unlike first world countries such as the United States (*Characteristics of Industry* 2013) or Canada, the composition of its sectors differ. This is because the more developed a nation is, the less focus is put into its primary and secondary sector and more of the population seek for jobs in the tertiary sector. On the other hand Malaysia; while still having a tertiary sector; relies much more on its primary and secondary sector for income. Due to this, the focus on vocational learning is much more encouraged than it would be in a first world country and the rate at which teachers look at their school options may differ.

We chose the approach of logistic regression as it fit our line of questioning the most. The choice was between two schools, so it was either they teach at an academic school or teach at a vocational one. We were able to use the data from Department of Statistics Malaysia (DOSM, Statistics Malaysia (2020)) and parse through the different categories to find the dataset needed. This data is collected via the Malaysian census. It was done through face to face interviews, self enumeration and e-census. After acquiring the data, we proceeded to clean it by only keeping the data needed. We focused on number of males and females and deleted rows with no information from these schools as they would not add to the statistic. Afterwards the type of school was chosen as our variable of interest and we proceeded to perform logistic regression. The reasoning and process is explained within the Data and Model section.

After using logistic regression and creating our model, we have found that overall teachers are more likely to choose teaching at an academic school over a vocational school. There are most definitely many external factors that effect this outcome such as income and frequency of schools. As teachers in Malaysia, there are

not many benefits, thus for many of these teachers' the deciding factors is their salary. The income for those teaching in an academic school is higher than those in vocational school (*Average Secondary School Teacher Salary in Malaysia* 2020).

Data

To explain in more detail about this data received from the DOSM, the method of self-enumeration would be when the government would drop off the census at homes and at a later date, collect the completed ones. As for the e-census, it would be for those with available internet connections where they would just verify their identity and perform an online survey.

The data we downloaded from the DOSM website was an excel file. Looking at this data, the 'School type' column stood out as it was either 'Academic' or 'Vocational College'. This could easily be a binary variable that could be turned into a logistic regression model, with sex and state being the independent variables. The data was shown by school and the frequency of male and female teachers at each school. The first step we took was to handle the non-responses. Some of the schools provided no number for male and female teachers, thus we eliminated those rows from the table.

Since we had decided that the school type would be our variable of interest, we would need to create a binary variable for that data. Our first problem was that the data provided frequency of males and females at each institution instead of individual responses. To fix this, we duplicated the rows equivalent to the frequency of sex given. If a row said 600 males, we would repeat said row another 599 times. Afterwards we deleted the frequency column. Then we created another column called 'bin_schooltype' in which if their 'School type' was 'Academic' they would be assigned the value 1 while 'Vocational College' was given 0.

After Cleaning the data, we proceeded into creating the Model.

Model

We created a general linear model to compute our coefficients for our logistic regression model. Since our 'State' variable is categorical, dummy variables were created that could be turned on and off for each category. We then created the model with our given coefficients and variables. Logistic regression has the following model:

$$log(\frac{p}{1-p}) = B_0 + B_1x_1 + ... + B_kx_k \tag{1}$$

Where p is the probability that the event of interest occuring (teaching at an academic school), B_0 is the y-intercept and $B_i, 1 \leq i \leq k$, coefficient represents change in log odds for one unit increase in x_i.

We then compute the generalised linear model with bin_schooltype as the dependent variable and sex and state being our independent variables.

We then called onto the summary(first_logit) function to retrieve all the needed coefficients and assigned simpler variable names to each value to form our regression formula:

- Intercept = 3.185e-01
- sexMale = -2.245e-14
- StateKedah = 4.925e-01
- StateKelantan = 1.924e-01
- StateMelaka = -3.185e-01
- StateNegeriSembilan = -1.361e-01
- StatePahang = 2.877e-01

- StatePerak = -1.178e-01
- StatePerlis = -3.185e-01
- StatePulauPinang = -3.185e-01
- StateSabah = 9.137e-01
- StateSarawak = 1.291e+00
- StateSelangor = 3.822e-02
- StateTerengganu = 3.747e-01
- StateW.P.KualaLumpur = -3.185e-01
- StateW.P.Labuan = -3.185e-01
- StateW.P.Putrajaya = 1.425e+01

Then after using equation (1):

$$\begin{gathered} log(\frac{p}{1-p}) = Intercept + sexMale(x_1) + StateKedah(x_2) + StateKelantan(x_3) + \\ StateMelaka(x_4) + StateNegeriSembilan(x_5) + StatePahang(x_6) + StatePerak(x_7) + \\ StatePerlis(x_8) + StatePulauPinang(x_9) + StateSabah(x_{10}) + StateSarawak(x_{11}) + \\ StateSelangor(x_{12}) + StateTerengganu(x_{13}) + StateW.P.KualaLumpur(x_{14}) + \\ StateW.P.Labuan(x_{15}) + StateW.P.Putrajaya(x_{16}) \end{gathered} \tag{2}$$

Then, when placing the coefficients in:

$$\begin{gathered} log(\frac{p}{1-p}) = 3.185e{-}01 - 2.245e{-}14x_1 + 4.925e{-}01x_2 + 1.924e{-}01x_3 - 3.185e{-}01x_4 - \\ 1.361e{-}01x_5 + 2.877e{-}01x_6 - 1.178e{-}01x_7 - 3.185e{-}01x_8 - 3.185e{-}01x_9 + 9.137e{-}01x_{10} + \\ 1.291e{+}00x_{11} + 3.822e{-}02x_{12} + 3.747e{-}01x_{13} - 3.185e{-}01x_{14} - 3.185e{-}01x_{15} + \\ 1.425e{+}01x_{16} \end{gathered} \tag{3}$$

When modeling the equations, x_1 is either on ($x_1 = 1$ for males) or off ($x_1 = 0$ for females). $x_2 - x_{16}$ are all dummy variables for states. This means that when one is active (the corresponding $x_i = 1$), the others are all 0.

Results

We can use the model created above to predict the statistic we have been looking for. We now just take the input of sex and state, place them accordingly into the equation and calculate the outcome.

Our first example could be a female from the state of Kedah. Using equation (2) to find the variable and equation (3) for the coefficients we get:

$$\begin{gathered} log(\frac{p}{1-p}) = Intercept + sexMale(0) + StateKedah(1) + StateKelantan(0) + \\ StateMelaka(0) + StateNegeriSembilan(0) + StatePahang(0) + StatePerak(0) + \\ StatePerlis(0) + StatePulauPinang(0) + StateSabah(0) + StateSarawak(0) + \\ StateSelangor(0) + StateTerengganu(0) + StateW.P.KualaLumpur(0) + \\ StateW.P.Labuan(0) + StateW.P.Putrajaya(0) \end{gathered} \tag{4}$$

$$log(\frac{p}{1-p}) = 3.185e{-}01 - 2.245e{-}14(0) + 4.925e{-}01(1) + 1.924e{-}01(0) - 3.185e{-}01(0) - 1.361e{-}01(0) + 2.877e{-}01(0) - 1.178e{-}01(0) - 3.185e{-}01(0) - 3.185e{-}01(0) + 9.137e{-}01(0) + 1.291e{+}00(0) + 3.822e{-}02(0) + 3.747e{-}01(0) - 3.185e{-}01(0) - 3.185e{-}01(0) + 1.425e{+}01(0) \quad (5)$$

We then get $p = 0.866$. This is the probability for a female from Kedah to be teaching in an academic school. If we continue the calculations for the rest of the possibilities we get the following table:

Table 1: Probability of Teachers Teaching at an Academic School

State	Female	Male
Kedah	0.866	0.866
Kelantan	0.764	0.764
Melaka	0.500	0.500
Negeri Sembilan	0.603	0.603
Pahang	0.802	0.802
Perak	0.614	0.614
Perlis	0.500	0.500
Pulau Pinang	0.500	0.500
Sabah	0.945	0.945
Sarawak	0.975	0.975
Selangor	0.695	0.695
Terengganu	0.831	0.831
W.P. Kuala Lumpur	0.500	0.500
W.P. Labuan	0.500	0.500
W.P. Putrajaya	0.743	0.743

And if we want to look at the opposite (probability of teaching at a vocational school), we get the following table:

Table 2: Probability of Teachers Teaching at a Vocational School

State	Female	Male
Kedah	0.134	0.134
Kelantan	0.236	0.236
Melaka	0.500	0.500
Negeri Sembilan	0.397	0.397
Pahang	0.198	0.198
Perak	0.386	0.386
Perlis	0.500	0.500
Pulau Pinang	0.500	0.500
Sabah	0.055	0.055
Sarawak	0.025	0.025
Selangor	0.305	0.305
Terengganu	0.169	0.169
W.P. Kuala Lumpur	0.500	0.500
W.P. Labuan	0.500	0.500
W.P. Putrajaya	0.257	0.257

An easier way to compare to put the values for the schools right next to each other. Since the values for males and females are identical, one table suffices:

Table 3: Probability of Teachers Teaching Either Academic or Vocational Schools

State	Academic	Vocational
Kedah	0.866	0.134
Kelantan	0.764	0.236
Melaka	0.500	0.500
Negeri Sembilan	0.603	0.397
Pahang	0.802	0.198
Perak	0.614	0.386
Perlis	0.500	0.500
Pulau Pinang	0.500	0.500
Sabah	0.945	0.055
Sarawak	0.975	0.025
Selangor	0.695	0.305
Terengganu	0.831	0.169
W.P. Kuala Lumpur	0.500	0.500
W.P. Labuan	0.500	0.500
W.P. Putrajaya	0.743	0.257

Discussion

From the results above, we some very strong evidence of a preference and surprise in the difference between sex. The most obvious result is how for majority of states, there is a higher probability for a teacher to teach at an academic school over a vocational one. The lowest the probability goes for favouring academic is 0.500 and even that does not show a preference, but rather an even split. There are several reasons for this. One reason could be that there are overall more students who go to an academic school, thus more teachers are needed to supplement these schools, leading to more teachers teaching there. Another reason could be that the skillsets for both these schools are vastly different and people may see a vocational school as needing much more specific knowledge. The pay difference is another factor that needs to be taken into consideration as on average teachers at an academic school earn more than those who teach at a vocational one.

One finding that surprised us was the lack of difference in sex values. In both tables 1 and 2, the females and males have identical probabilities. While these values are not exactly identical, the difference between them is too small to matter. As when looking at equation (2), sex is x_1 and has a coefficient $-2.245E{-}14$. Since this is a dummy variable, we let $ x_1 = 1 or 0 $. This is such a small value that it does not affect any significant figures. This is what ultimately leads to no difference in values. As for why the coefficient was so small; it could be due to the values of males to females recorded from vocational and academic schools being identical, making sex a weak indicator.

So how can we apply this information for the future? This all depends on where the leaders of Malaysia want to take its country. If they want more focus in the secondary sector, this research would tell them to invest more into the teachers for vocational school and create more incentives for these jobs. If the leaders are leaning towards the tertiary sector, this research would affirm their actions as there are more teachers in academic schools which lead to more jobs in the tertiary field.

Weaknesses and Future Work

As for weaknesses in this research, it could be within the linearity of its independent variables. Because after the calculations, it seems like sex was a useless variable to choose as an indicator. Maybe more factors could

have been taken into consideration, such as age and house address. Just anything that would be a stronger indicator as this regression mostly based on states.

For the future of this research, it could be extended into the next census and more specific questions could be asked. Maybe some open questions to get a better scope of the peoples' opinions.

Note

```
# Code and data supporting this analysis is available at: https://github.com/fadlan-arif/malaysia-censu
```

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *Rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown.

Average Secondary School Teacher Salary in Malaysia. 2020. PayScale. https://www.payscale.com/research/MY/Job=Secondary_School_Teacher/Salary.

Characteristics of Industry. 2013. BBC. https://www.bbc.co.uk/bitesize/guides/zx3vtyc/revision/2.

Malaysia: Distribution of Employment by Economic Sector from 2010 to 2020. 2020. statista. https://www.statista.com/statistics/319036/employment-by-economic-sector-in-malaysia/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Statistics Malaysia, Department of. 2020. *Open Data*. DOSM. https://www.dosm.gov.my/v1/index.php?r=column3/accordion&menu_id=amZNeW9vTXRydTFwTXAxSmdDL1J4dz09.

Ushey, Kevin, JJ Allaire, Hadley Wickham, and Gary Ritchie. 2020. *Rstudioapi: Safely Access the Rstudio Api*. https://CRAN.R-project.org/package=rstudioapi.

Vocational School. 2020. Wikipedia. https://en.wikipedia.org/wiki/Vocational_school#References.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2019. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.