QMSS Data Analysis Final Project

The Effect of Working Fields/Industries on Foreign Workers' Job Stability

Sile Yang (UNI: sy2738)

Introduction

Many foreign nationals work in America nowadays. However, it is tough for an international worker to keep job stable in America. They are faced with too many challenges including strict government policies, slumping economy, serious competition in the company, and so on. Based on online discussion and personal experience, it seems that foreigners' job stability depends on their working fields. For example, an international worker who works as a software developer might be more stable than one who works in the digital media field. Therefore, in this research, I will test whether working in different fields/industries might affect an international worker's job stability.

I will use the H-1B database from the United States Department of Labor. My independent variable is *working fields* such as computer science, financial, physicians, management, and so on. In the following part, I will discuss how I clean the data to get the independent variable. The dependent variable is *job stability*, which is reflected by H-1B status.

In my expectation, for foreign workers, working in different fields might affect their job stability. Particularly, foreign workers who work in the computer science field might have higher job stability.

Hypotheses

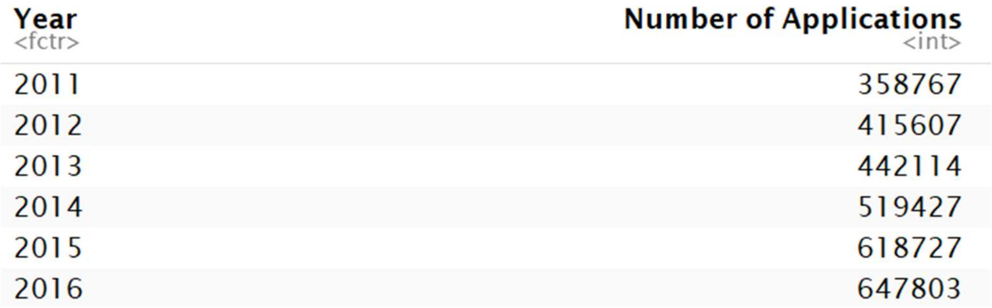
1. Working in different fields affects a foreign worker's job stability.

2. For international workers, working in the computer science field comes with a higher job security than other fields.

Description of Dataset and Variables

The dataset is stem from the United States Department of Labor. The Office of Foreign Labor Certification releases H-1B data annually, and users have access to download disclosure data in OFLC official website. The H-1B program allows foreign workers to be temporarily employed by employers in the U.S. on a nonimmigrant basis in specialty occupations or as fashion models of distinguished merit and ability. A specialty occupation requires the theoretical and practical application of a body of specialized knowledge and a bachelor's degree or the equivalent in the specific specialty (e.g. sciences, medicine, healthcare, education, biotechnology, and business specialties, etc.). Current laws limit the annual number of qualifying foreign workers who may be issued a visa or otherwise be provided H-1B status to 65,000 with an additional 20,000 under the H-1B advanced degree exemption. The data, which contains 3,002,458 observations, is collected officially by each H-1b application from 2011 to 2016.

Table.1: The number of applications in 2011 to 2016



Year	Number of Applications
<fctr>	<int>
2011	358767
2012	415607
2013	442114
2014	519427
2015	618727
2016	647803

The raw data is messy, so I did data cleaning to ensure my further analysis.

Firstly, my independent variable is *working fields*. I created this variable by using *SOC_NAME* in raw dataset. The original variable shows respondents' job (like computer systems analyst, computer programmers, financial analyst, biological scientists, and so on). I classified such jobs and created *working fields*, and categorized them under computer and non-computer. In this dummy variable, 0 represents a worker does not work in computer-related fields while 1 represents a worker work in the computer field. Table 2 shows the number of applications in each field and their coding numbers.

Table.2: Working Fields and Number of Applications

Working Fields	*Number of Applications*	*Coding Number*
Computer	1,777,383	1
Non-Compute	759,142	0

Table.3: Job Stability and Number of Applications

Working Fields	*Number of Applications*	*Coding Number*
Stable(CERTIFIED)	2,226,065	1
Unstable (OTHER STATUS)	310,460	0

Secondly, my dependent variable is *job stability*, which is reflected by H-1B status. Job stability means that one employee works in a company for a long time without any unemployment caused by various reasons. In the dataset, the variable *CASE_STATUS* shows several statuses of H-1B: Certified, Certified – Withdrawn, Withdrawn, Denied, Rejected, and others. Except for "Certified" status; other statuses indicate that the foreign worker has left the company, or the workers have been terminated; so, only "Certified" status shows the job stability of foreign workers.

I made 1 to represent Certified H-1B Status, and 0 represents others, where 1 indicates good job stability. Finally, after data cleaning, the total population is 2,536,525.

Table.4: Mean, Standard Deviation, and Range,

Variable	*Obs*	*Mean*	*Std.Dev*	*Min*	*Max*
Stability	2,536,525	0.878	-	0	1
Workfields	2,536,525	0.701	-	0	1

Table.4 shows descriptive statistics of variables. The average of *stability* shows that 0.878 is the fraction of stable job status. Similarly, the mean of *Workfields* shows that about 70 percent of respondents work in computer science related fields. Because both variables are dummy variables, the standard deviation is meaningless in this case.

Initial Model

1. Linear Probability Model

I will start with the simplest model. The initial model of the association between working fields and job stability adopts the linear probability model (both working fields and job stability are dummy variables). With a dummy variable as the dependent variable, the simple model can be called a linear probability model. Table.5 shows the results of the model.

Table.5: The results of LPM (stability ~ working fields) (Initial Model)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
Constant	0.840	0.0003	2240.6	0.000 ***
Working Fields	0.053	0.0004	118.1	0.000***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

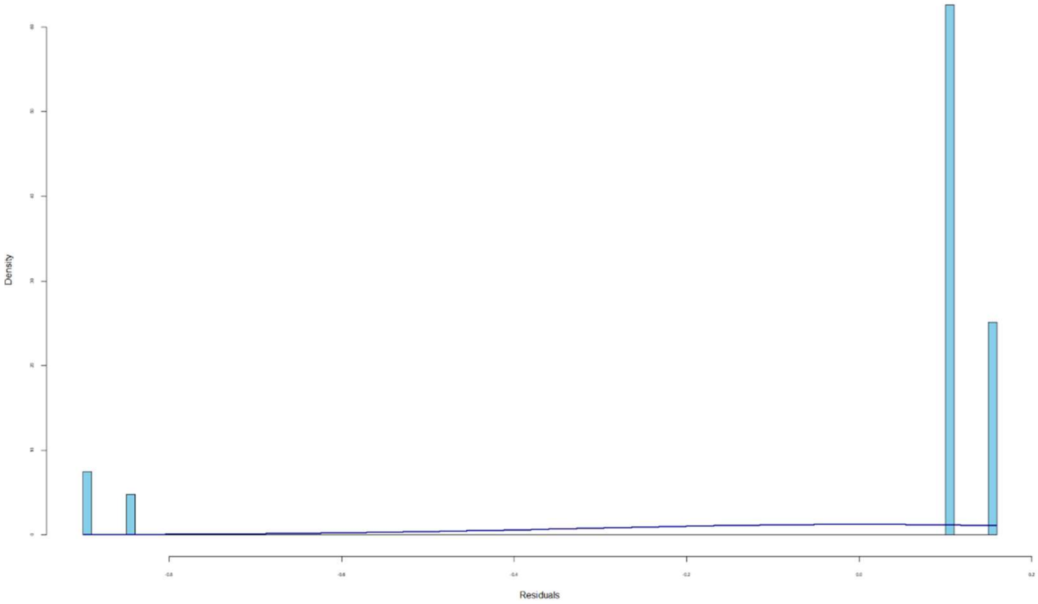
In table.5, the constant shows that a worker who does not work in a computer-related field is 84.3 percent points more likely to work in a stable condition, on average. Compared with people who do not work in computer-related fields, people working in computer-related fields are 5.3 percent points more likely to work in a stable condition. The results of the estimate are not as same as what I expected. Interestingly, the coefficient of the variable *working fields* is too small, even though the t-value is statistically significant.

2. Remarks on the Linear Probability Model

In this part, I will test the regression. Unfortunately, the results show that the initial model is insufficient. There are several limitations of this model.

Firstly, as figure.1 showing, in the initial model, the normality assumption of errors and residuals is violated. The normality assumption of errors is a basic assumption of linear regression. The residuals are not in normality distribution. Therefore, the p-value of the initial model is questionable.

Figure.1 The distribution of residuals



Secondly, I test the homoscedasticity by Breush-Pagan test and the NCV test. The homoscedasticity is an assumption that the variance around the regression is the same for all values of the independent variable. Both values of the test are 0 (less than a significance level of 0.05). Thus we should reject the null hypothesis that the variance is same and infer that heteroscedasticity is present.

Table. 6 the results of BP test and NCV test

	Estimate	*Df*	*p-value*
Breush-Pagan test	BP = 13330	1	0.00***
NCV test	Chisquare= 35177.94	1	0.00***

Besides failing to pass the linear regression test, the model has other limitations. Firstly, the initial model only contains one independent variable. However, there might be other omitted variables and alternative explanations. For instance, does a foreigner's wage affect his or her stability? Is there any possibility that the year that a foreign worker is hired might influence his or her job stability? We should add more controlling variables into the model to make it more fitted. Secondly, the categories of the independent variables are too simple. Non-computer fields might be so overly broad that we ignore variances in other fields. For example, the variance of job stability in computer fields and medical fields might differ from the variance in computer fields and education fields. Therefore, we should specify the working fields in the following part.

Model Improvement

1. Specify working fields

Based on the previous independent variable, non-computer fields have been manually classified into eight fields: Computer Science (including Data Science), Management, Business, Engineering, Medical, Law, Biology, Chemistry, Media and Art, and Education. Table .7 shows the number of applications in each field and their coding numbers. I made *Working fields* into eight dummy variables and ran a linear probability model.

Table.7 Working Fields and Number of Applications

Working Fields	*Number of Applications*	*Coding Number*
Computer Science	1777383	0
Management	95231	1
Business	193174	2
Engineering	173777	3
Medical	148541	4
Law	8055	5
Biology	54105	6
Chemistry	20947	7
Media and Art	38736	8
Education	26576	9

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Table.8 shows the results of model 2. The direction of coefficients is as I expected. Besides constant, they are all negative, confirming that compared to computer-related fields, other fields are less likely to get a stable job. The results support my expectation that compared to other fields, a foreign employee who works in a computer-related field is more likely to get job stability, on average. However, this model still cannot explain some alternative explanations. Thus, we should add control variables to it.

Table.8 the results of LPM (stability ~ working fields) (Model 2)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
Constant	0.893	0.0002	3646.79	0.00***
Workingfields.Management	-0.021	0.0010	-19.38	0.00***
Workingfields.Business	-0.049	0.0008	-62.88	0.00***
Workingfields.Engineering	-0.042	0.0008	-51.05	0.00***
Workingfields.Medical	-0.060	0.0009	-68.10	0.00***
Workingfields.Law	-0.092	0.0036	-25.29	0.00***
Workingfields.Biology	-0.113	0.0014	-79.25	0.00***
Workingfields.Chemistry	-0.075	0.0022	-33.23	0.00***
Workingfields.Media&Art	-0.068	0.0017	-40.75	0.00***
Workingfields.Education	-0.052	0.0020	-25.77	0.00***

2. Add control variables

Based on the original dataset, there are three controlling variables:

Wage: the application's current wage.

Job Status: Full time (1) or part time (0). In the raw data, there is one variable *FULL_TIME_POSITION* showing the job status. I made 1 to represent full time job status, and 0 is part time job status.

Year: the year in which the worker applies H1-B.

Table.9 Mean, Standard Deviation, and Range,

Wage	*2,536,525*	*149,505.9*	*5,602,007*	*0*	*6,997,606,720*
Year	2,536,525	2014	1.661	2011	2016
Status	2,536,525	0.871	-	0	1

Table.9 shows descriptive statistics of control variables. The mean of *wage* is 149,505.9, ranging from 0 to 6,997,606,720. It is a wide gap for a wage with 5,602,007 standard deviation. Then, I calculated the kurtosis of *wage*, which is 960,929; far from the value of the normal distribution (A normal distribution has a kurtosis of 3). The average of *year* is 2014, which means that we have more data in 2015 or 2016. The standard deviation of *year* is not very large. As for status, the average indicates that about 87.1 percent of respondents have a full-time job.

The descriptive statistics show that the *wage* is not a good variable. To get a better model, to make *wage* more normal and increase interpretability, I made a log transformation of *wage*. I created a new variable *lnwage*, and figure.2 shows the distribution of this new variable.

Figure.2: the distribution of *lnwage*

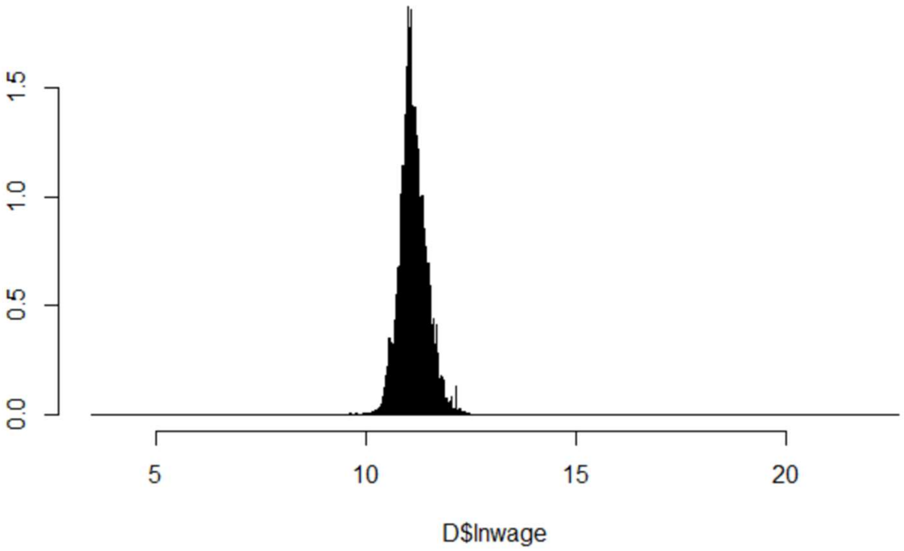


Table 10 shows the results of the linear probability model when the control variables are included (Model 3). Same as model 2; the coefficients of *working fields* are negative, inferring that compared with computer-related fields, a foreign worker on average is less likely to get a stable job in other fields, controlling other variables. However, after adding control variables, the coefficients become slightly high (except for *biology*). As for control variables, the results show: 1) net of other variables, a 1 percent increase in wage leads to 3.1 percent points less likely to work in a stable condition; 2) net of other variables, with one year increase, a worker is 0.5 percent more likely to work stably; 3) net of other variables, compared with part time job, the full-time workers are 2.4 percent more likely to get a job stable. As for constant, for workers who work in computer-related fields, with no wage and part-time job in 0 years, they are 98.21 percent less likely to get a stable job.

However, the model 3 still cannot eliminate the limitations I mentioned above. The linear regression model might not be an appropriate statistical method to interpret discontinuous variables.

Table.10: the results of LPM (stability ~ working fields + control variables) (Model 3)

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
Constant	-9.821	0.2770	-35.45	0.00***
Workingfields.Management	-0.018	0.0011	-16.70	0.00***
Workingfields.Business	-0.048	0.0008	-60.17	0.00***
Workingfields.Engineering	-0.039	0.0008	-47.08	0.00***
Workingfields.Medical	-0.054	0.0009	-61.44	0.00***
Workingfields.Law	-0.077	0.0036	-21.08	0.00***
Workingfields.Biology	-0.123	0.0014	-84.97	0.00***
Workingfields.Chemistry	-0.077	0.0023	-33.89	0.00***
Workingfields.Media&Art	-0.071	0.0017	-42.12	0.00***
Workingfields.Education	-0.060	0.0020	-29.54	0.00***
Lnwage	-0.031	0.0006	-50.78	0.00***

Year	0.005	0.0001	39.72	0.00***
Status	0.024	0.0006	34.50	0.00***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

3. Logistic Regression Model

Considering the limitations of the linear regression models above; I adopt a logistic regression model to analyze. The logistic regression model is a model that predicts the log-odds linearly.

There is a prediction model:

$$\text{Logit(stability)} = \beta0 + \beta1 * \text{Workingfields.Management} + \beta2 * \text{Workingfields.Business} + \beta3 * \text{Workingfields.Engineering} + \beta4 * \text{Workingfields.Medical} + \beta5 * \text{Workingfields.Law} + \beta6 * \text{Workingfields.Biology} + \beta7 * \text{Workingfields.Chemistry} + \beta8 * \text{Workingfields.Media\&Art} + \beta9 * \text{Workingfields.Education} + \beta10 * \text{Lnwage} + \beta11 * \text{year} + \beta12 * \text{status} + \mu$$

The regression coefficient for each independent variable measures the effect of a one-unit change in that variable on the log-odds of the dependent variable. The null hypothesis of this model is that all β is zero. In this part, we will test this hypothesis.

The table.11 shows the results of the logistic regression model (model 4). The p-value shows that all coefficients are statistically significant, which means that we should reject the null hypothesis. The direction of all coefficients is the same as that of the above model, confirming that compared with computer-related fields, and working in other fields decreases their chances of working in stable condition, net of other variables. Particularly, controlling other variables, for a foreigner, working in business fields decreases its logit by 0.048, compared with computer fields. Similarly, for a foreigner, net of other variables, working in art and design fields might decrease its logit by 0.071, compared with people who work in computer-related fields.

Regarding odds ratios, 1) net of other variables, compared to people who work in computer fields, being a foreigner increases the odds of getting a stable job by less than 1, which indicates that such foreigners are less likely to get a stable job compared to an international worker in computer fields. For instance, compared to computer science workers, the odds of getting a stable job for a foreigner who works in business fields go up by –34.1% ((0.659 - 1) * 100%). The odds for people who work in media and art fields go up by -44.5%((0.555 - 1) * 100%), and the odds for people in education fields increase by –39.9% ((0.601 - 1) * 100%). 2) net of other variables, with one percent increase in wage, the odds of getting a stable job go down by 20.2% ((1 – 0.798) * 100), on average. 3) net of other variables, with one-year increases, the odds of job stability go up by 4.9% ((1.049 - 1) * 100%), for foreign workers, on average. 4) net of other variables, compared with part-time workers; for a full-time foreign worker, the odds of getting a stable job go up by 22.1% ((1.221 - 1) * 100%).

Table.11 the results of Logistic Regression Model (Model 4)

	Odds Ratios	*Estimate*	*Std. Error*	*z value*	*Pr(>|t|)*
Constant	0.000	-91.314	0.2770	-35.81	0.00***
Workingfields.Management	0.836	-0.179	0.0011	-17.87	0.00***
Workingfields.Business	0.659	-0.417	0.0008	-61.00	0.00***
Workingfields.Engineering	0.702	-0.354	0.0008	-49.24	0.00***
Workingfields.Medical	0.627	-0.467	0.0009	-62.61	0.00***
Workingfields.Law	0.541	-0.615	0.0036	-21.83	0.00***
Workingfields.Biology	0.395	-0.929	0.0014	-85.30	0.00***
Workingfields.Chemistry	0.532	-0.630	0.0023	-34.79	0.00***

Workingfields.Media&Art	0.555	-0.588	0.0017	-43.09	0.00***
Workingfields.Education	0.601	-0.510	0.0020	-29.76	0.00***
Lnwage	0.798	-0.226	0.0006	-47.63	0.00***
Year	1.049	0.048	0.0001	37.50	0.00***
Status	1.221	0.200	0.0006	32.02	0.00***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

4. Results and discussion

In this part, I will discuss the best model and the limitations of this project. As discussed above, I created four models; three of which are linear probability models, and the last is a logistic regression model. Table .12 shows the comparison between the four models. As table.12 showing, for model 1,2,3, the adjusted R-square increases, confirming that the improvement works, and model 3 is better than model 1 and 2. I also use a partial F-test to test whether the variables improve the models fit. The F-value of model 2 and 3 indicate that adding variables into models improves the fit. All of the results support that model 3 is the best model among model 1,2,3. The next step is to evaluate model 4. Instead of using R-square, I use a chi-square test to indicate how well the model 4 fits the data. The value of chi-square is 499.2. With statistical significance, it supports that the model 4 fits the H1-b dataset.

Comparing model 3 and model 4, I prefer to select model 4 as a final model. Because most of my variables are not continuous; indicating that the simple linear regression model might not be the right fit for the data. However, logistic regression deals with this problem by using a logarithmic transformation on the outcome variable which allows me to model a nonlinear association linearly. Besides, I can interpret variables better in the logistic regression model.

Table.12: the comparison of models

	Model 1	*Model 2*	*Model 3*	*Model 4*
Constant	0.840***	0.893***	-9.821***	-91.314***
Working fields	0.053***	-	-	-
Workingfields.Management	-	-0.021***	-0.018***	-0.179***
Workingfields.Business	-	-0.049***	-0.048***	-0.417***
Workingfields.Engineering	-	-0.042***	-0.039***	-0.354***
Workingfields.Medical	-	-0.060***	-0.054***	-0.467***
Workingfields.Law	-	-0.092***	-0.077***	-0.615***
Workingfields.Biology	-	-0.113***	-0.123***	-0.929***
Workingfields.Chemistry	-	-0.075***	-0.077***	-0.630***
Workingfields.Media&Art	-	-0.068***	-0.071***	-0.588***
Workingfields.Education	-	-0.052***	-0.060***	-0.510***
Lnwage	-	-	-0.031***	-0.226***
Year	-	-	0.005***	0.048***
Status	-	-	0.024***	0.200***
Adjusted R-squared	0.0055	0.0068	0.0082	-
F - value	-	417***	1237***	-
Chi - square	-	-	-	499.2***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

However, although I think model 4 is the best model among all, it still has several limitations. Firstly, the model 4 still cannot pass the heteroskedasticity test with a value less than a significance level of 0.05. Secondly, this limitation is stem from the original dataset, which does not contain sufficient variables to analyze an issue. The raw H-1B dataset cannot provide appropriate variables to analyze more alternative explanations. For instance, it is possible that gender, race, and age might affect a foreigner's job stability, and these are not included in the original dataset. Due to time constraint, I cannot find another data resources or combine data to improve my model.

Conclusion

The final model supports my initial hypothesis and confirms that working in different fields affects a foreigner's job stability, on average, with controlling wage, starting year and job status. In particular, in the same condition, an international worker who works in computer fields is more likely to get a stable job, on average.

The research still needs further study. Firstly, as I mentioned above, some basic but important basic personal information is not included in the original dataset. I need to find more related data resources, and, if necessary, to combine several datasets. Secondly, the theoretical fundamental of job stability is weak. Study of more scholarly articles is required to find a solid way to measure job stability.