Stat 433 Final Project Report

Taran Katta, Nolan Peterson, Mark Wu

12/17/2021

Introduction

The aim of this project is to analyze the relationship between age in occupational sectors and labor shortages within occupational sectors in Wisconsin. Furthermore we wish to test how well labor shortages can be predicted using ages within said occupational sectors. The motivation behind exploring this relationship is that there is a commonly held idea that jobs which are held mostly by older workers are more prone to shortages. For instance, the average age of farm workers in Wisconsin is 53, which is considerably greater than the average age of all workers in Wisconsin. Additionally there is a shortage of farm workers in Wisconsin. It would not be illogical to assume that there is a connection between these two facts. In this report we will explore this question. We will examine ages within occupations in 2016 and examine shortages within those same occupations in 2019. This relationship is visualized in figure 1.

Figure 1: Theorized relationship between Age and Shortage

Thesis Statments

In this paper we argue that the average age within occupational sectors has bearing on the probability of labor shortage in that occupational sector.

Data

The data for this project come from two primary sources. The first is American Community Survey (ACS) Person data for Wisconsin. And the second is the Occupational Employment and Wage Statistics (OEWS) state data for Wisconsin also. (Add reference for link) The data in its raw form required a lot of reprocessing. The ACS data for 2016 and 2019 together was 118,336 observations with 288 variables. The vast majority of these variables were not useful to answering our research question. Consequently only the following variables were selected from the ACS data: OCCP (Occupational Code), AGEP (Age of the Person), PWGTP (Replicate Weight Estimate) and ESR (Employment Status). This subset of data was then grouped by year and occupation code, so that a weighted estimate could be constructed of the mean age within occupations and unemployment rate within occupations. The result was data set with 4 variables and 996 observations representing approximately 498 occupations across two years (2016 and 2019).

This grouped data frame was joined to the OEWS data for Wisconsin for 2016 and 2019 by occupation code. Because there are fewer ACS occupational codes than there are OWES occupational codes, this means that ACS codes tend to be broader and represent more occupations that OWES code. Using a crosswalk key between these two code systems it was determined which codes mapped to each other. In instances where there was multiple OEWS codes for one ACS code, the OEWS observations were merged into one observation. Additionally there were occupations were the sample sized collected was too small to estimate unemployment rates within that occupational sector. These occupations were dropped from the data set, for reasons that will be explained in more detail in the methods section. With this cleaned and joined data estimates of job growth rate and wage growth rate could be constructed. These were added to the data set as additional variables. And finally the data was pivoted wider by year so there was individual variables for both 2016 and 2019. The 79 observation was dropped from the data because it was a very extreme outlier in Job growth rate. The final data set has 83 observations and 22 variables. The steps for the data filtering cleaning and joining are summarized in figure 2.

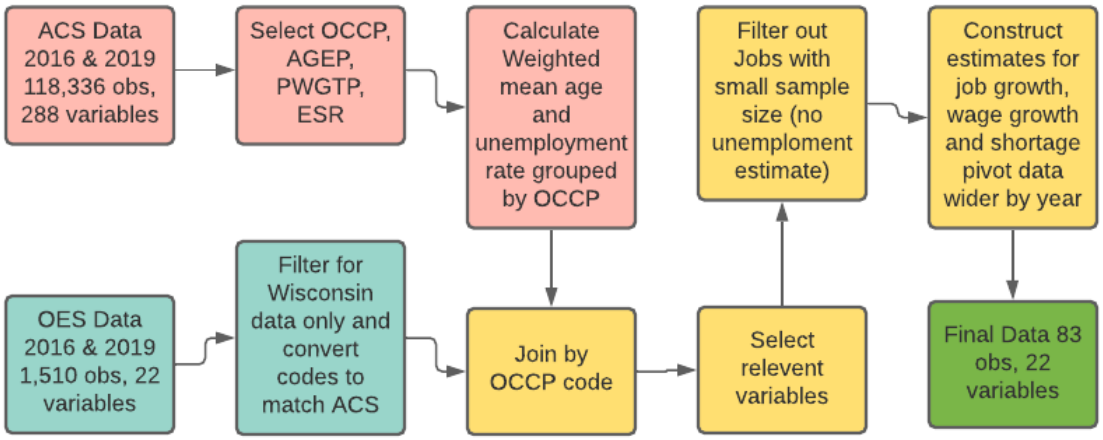

Figure 2: Data Pre-processing Summarized

Methods

Defining a Labor Shortage

A primary challenge of this project was identifying labor shortages in the 2019 data. A labor shortage has no exact definition in economics. However there is basis for defining a shortage within a given occupational sector with the following criteria (Barnow, 2016) and (Veneri, 1999).

An occupational shortage is defined by three criteria:

1. The occupation's employment growth rate is faster than average employment growth
2. The wage increase is at faster than average.
3. The occupation's unemployment rate is below average.

In order to define a shortage, it was necessary to first construct estimates of the employment growth rate and wage growth rate. As mentioned in the data section of this report, these estimates were calculated from OEWS data from 2016 and 2019. The formulas used to do so can be found below.

$$Average \;\; Annual \;\; Employment \;\; Growth = \frac{Employment_{2019}}{Employment_{2016}} \times \frac{1}{3 \; Years - 1} \times 100\%$$

$$Average \;\; Annual \;\; Wage \;\; Growth = \frac{Avg \; Salary_{2019}}{Avg \; Salary_{2016}} \times \frac{1}{3 \; Years - 1} \times 100\%$$

The unemployment rate used is the unemployment in 2019.

With these criteria we were able to identify 12 occupations with shortages and 71 without shortages.

T-test for difference in Sample Means

With the labor shortage properly defined we sought to test the following hypothesis:

H_0 : Mean of all occupation age averages in shortage group = Mean of all occupation age averages in non-shortage group

H_a : Mean of all occupation age averages in shortage group > Mean of all occupation age averages in non-shortage group

This can be visually explained by looking at the density plots of the ages for these two groups in figure 3.

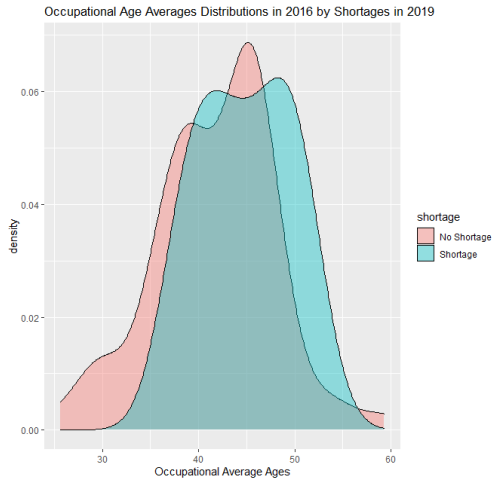

Figure 3: Density plots of age for shortage and non-shortage

An appropriate method to test hypothesis is a t test for difference of sample means. The t test statistic is calculated using the below formula.

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n_2}}}$$

In our case the t-test statistic was calculated as such:

$$t_{obs} = \frac{44.80 - 41.75}{\sqrt{\frac{4.79^2}{12} + \frac{6.19^2}{71}}} = 1.95$$

This value is then compared to the upper-tail critical from the student's t-distribution. Which in this case is 1.66. This comparison yields a p-value of 0.027. Which would lead us to reject the null hypothesis at the $\alpha = 0.05$ level.

However, a 95% confidence interval can be calculated for the difference of sample means using t_{obs}. It has the following form:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{obs} \times S_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where S_p is the pooled sample standard deviation and has the form:

$$S_p = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$$

When this was applied to our data this was the result:

$$(44.80 - 41.75) \pm 1.95 \times 6.02 \times \sqrt{\frac{1}{12} + \frac{1}{71}} = (-0.61,\ 6.71)$$

As can be seen in the above interval for difference of means 0 is at the far end of the interval but not excluded from it.

Logistic Regression Model

In the data set

References

Appendix