EDA Project

Cristen McCann, Yilin Song, and Vitaly Ievlev

11/15/2017

### **Introduction**

In this project we are exploring statistical relationships between depression and different socio-economic factors that might contribute to it. In order to accomplish this task, we examined the data that was collected as a part of Job Intervention Program evaluation that was conducted in the University of Michigan in 1992. The data were originally collected from 1801 unemployed adult respondents and later organized into a table with 899 observations of 17 different variables. The original study focused on effectiveness of the job search training program, therefore the original dataset included variables before and after the training (treatment). Examples of these variables are depression, economic hardship and other ones (addressed later in the paper). We decided to take a different approach to analysing these data by focusing only on the socio-economic aspects and depression before treatment. Therefore we deleted certain variables from the dataset.

### **Variables Deleted from Original Dataset:**

Because we decided to focus socio economic factors that lead to depression in general rather than the outcomes of a specific job training program, which was the original intention of this dataset. We deleted the variables “treat”, “job\_seek”, “depress\_2”, “work1”, “comply”, “control”, “job\_dich”, and “job\_disc”, because they had to do with the execution and aftermath of the job training program, which is irrelevant to our question of study. We do not have to remove any observations, because they all have data for every variable. Removing the data above will not likely have an affect on our conclusion or have any implications, because we can’t generalize the conclusions anyway. We can only make conclusions about depression for unemployed people within the study, as the study is not randomly sampled. Therefore, data about newly employed people is irrelevant.

### **Variable Descriptions**

The depression variable is continuous quantitative, with values from 1-5. This information was gathered by asking each participant 11 questions about their own experiences with depression, for which they answered on a categorical scale from 1-5 (meaning “not at all” to “extremely”) and the answers to the 11 questions were averaged. This method is based on the Hopkins Symptoms Checklist, which had 25 questions about overall mental health, with answers ranging from 1-4, which mean “Not at all,” “A little,” “Quite a bit,” and “Extremely” respectively (the depression score was calculated by averaging answers for 15 questions about depression specifically). The economic hardship variable is discrete quantitative, for which the study took the mean of 3 questions with answers from 1-5 (5 indicated highest economic hardship level). A survey was used to gather demographic information of age, sex, education, marital status, occupation, family income, and ethnic/racial identification. The age variable is continuous quantitative. The sex variable is categorical, with the 2 categories (male and female). The education variable is categorical with 5 categories (less than high school, high school, community college, bachelors, and graduate). The marital status variable is categorical, with 5 categories (married, never married, separated, divorced, and widowed). The occupation variable is categorical with 7 categories (professionals, operatives, managerial, sales workers, craftsmen, clerical, and laborers/service). The family income variable is categorical, with 5 categories (50k+,40-49k,25-39k, 15-24k, less than 15k). The ethnic/racial variable is categorical, with 2 categories (white and non-white). We didn’t do any data manipulations.

### **Single Variable EDA:**

#### **1. Depression**

We did the statistics summary for our depression status first and found out that the mean score of depression in this sample is 1.87 with a 0.566 standard deviation (Table.1). According to the histogram of depression score as a single variable, the distribution is approximately bell-curved, with a marked right skew (Fig.1).

#### **2. Economic Hardship**

The mean level of economic hardship is 3.02 on a scale of 1-5, and the standard deviation is 0.99 (Table.2)

#### **3. Age**

According to the statistics summary of age, the average age of the participants in the study is 37.56 with a 10.45 standard deviation, which means the age of the participants varies in a large range (Table.3). The histogram of age seems very normally distributed, but with a slight right skew (Fig.2).

#### **4. Sex**

There are 65 more females than males in the study (Table.4).

#### **5. Marital Status**

Most of the participants in our study are either married (45%) or never married (31%), followed by divorced (18%), separated (3%) and widowed (2%) (Table.5).

#### **6. Race**

There are much more whites (83%) included in this research than non-whites (Table.6).

#### **7. Education:**

Most of participants get at least a high school degree (Table.7).

#### **8. Income:**

The proportion of people in the 5 income categories are about the same (22%) except the group with 40 to 49 thousand dollars which is lower (12%) (Table.8).

### **Two Variable EDA:**

In our exploratory data analysis (EDA) we are examining the following relationships:

#### **1. Sex and depression**

For a comparison of the variables sex and depression, the mean depression score is 1.92 for women, and 1.82 for men. The standard deviation is 0.55 for men and 0.58 for women. The boxplot shows a slightly larger depression score median and range for females than males (Fig.3). The histogram shows a more normal distribution for females than males. The male histogram has a slight right skewed. We should explore whether females are significantly more depressed.

#### **2. Race and depression**

When comparing race and depression, the mean depression score is 1.80 for non-white, and 1.88 for white subjects. The standard deviation is 0.57 for non-white, and 0.56 for white subjects. The boxplot shows a much larger depression score median for whites (since white sample size is much larger) and but the ranges for the groups are similar (Fig.4). The histogram shows a more normal distribution for whites than non-whites. The non-white histogram has a dramatic right skew. We should explore whether whites are significantly more depressed.

#### **3. Income and depression**

When comparing income (5 groups) and depression, the mean depression score and standard deviations are largest for the middle group, the 40-49k (mean = 1.92, sd = 0.50), and also larger than the rest for the lowest and highest groups, creating a wavelike distribution. The boxplot shows a larger range for the 40 to 49k group, and lower median for the more than 50k group (Fig.5). Since the depression score in the 50k group is lower than the other groups, we want to do ANOVA to test whether there is a difference in depression scores between different income groups.

#### **4. Education and depression**

Comparison of education level and depression using a boxplot yielded some interesting statistics (Fig.6). For example, it was clearly evident that people who didn’t finish high school were the most depressed on average with a mean depression score at 2. While people who received high school, bachelors and professional degrees all have the same mean depression at approximately 1.84. Hence, we want to do ANOVA to test the relationship. However, the p-value is 0.313 which is higher than 0.05. We fail to reject there is no difference in depression scores between groups.

#### **5. Marital status and depression**

Assessment of the boxplot of the marital status and depression showed that married people had a lower mean of depression than all the other groups (Fig.7). Mean depression score for married people was 1.81, while the highest means were among divorced and widowed participants (means 1.94 and 1.97 respectively). We will expect to see there are some difference betwee different marital status. However, the ANOVA shows us the p-value is 0.0854 which is bigger than 0.05, so we fail to reject Ho.

#### **6. Occupation and depression**

Boxplot of depression means for different occupations shows us that clerks and sales workers seem to be less depressed on average with the means at 1.82 and 1.83 respectively. While the highest level of depression is observed among the laborers and service workers (mean 1.94) (Table.9). However, since the differences are not obvious, we don’t think there is a relationship between occupation and depression as we expected.

#### **7. Age and depression**

According to our scatter plot of depression score vs. age, the dots are spread out evenly. Therefore, we can’t find any association between depression scores and age based on our scatter plot. After we did the xy plot for depression score and economic hardship, there is no clear association between them either, since the points are also somewhat-evenly spread out on the graph (Fig.8). However, we would like to do future regressions to figure out whether there are some relationships between them.

#### **8. Economic hardship and depression**

We represented the relationship between economic hardship and depression in a form of a box plot rather than a scatter plot because it conveys the relationship better. By looking at the plot it is clear that there is an association between the two variables (p<0.0001) (Fig. 9). People who experience economic hardship are clearly more depressed. For example among those with least economic hardship mean depression score was 1.5, while the mean depression score for those in the highest economic hardship group is 2.12 (Table.10).

Considering that some of the variables addressed above might be influencing depression transiently through their influence on economic hardship, we also examined the following relationships:

#### **9. Marital and econ hardship**

We did both the histogram and boxplot (Fig. 10). for marital status and economic hardship. We found out that the mean economic hardship for widowed, separated, married and never married groups is about the same, but the mean economic hardship for divorced group is higher. The distributions of married and never married groups are more symmetric and normal compared to the other three (because the last three categories are also the smallest). Also, the IQR in the married group have the lowest economic hardship scores (highest, not largest), while the divorced and separated groups have the highest IQR. However, according to the statistics summaries, the means in different groups don’t differ too much from one another, and the standard deviations are all about 1. Therefore, we still want to do future analysis to test whether marital status influences economic hardship.

#### **10. Ocup and econ hardship**

The analysis of relationships between economic hardship and occupation is represented in Fig 11. For all of the occupations except two the level of economic hardship was at 3, but clerks had the lower mean economic hardship at 2.88, while craftsmen experience on average more economic hardship with the mean value of 3.15 (Table.11). Since the average scores don’t vary a lot. We expect there is no relationship betwee occupation and economic hardship.

#### **11. Race and econ hardship**

After analysis of race and economic hardship we found out that white respondents experience significantly (p=0.0007) less economic hardship than non-white respondents (Fig. 12). The mean economic hardship for white respondents is 2.97 while for non-white respondents the mean value is 3.29 (Table 12). Although this relationship is not extremely surprising, the fact that white people still experience more depression while having less economic hardship is intriguing. Especially considering the very strong positive correlation (p<0.0001 Fig. 9) between the economic hardship and depression.

### **Analysis Plans**

We already analyzed each of our variable pairs using t-tests, ANOVA, and chi-squared tests, in order to quantify the statistical significance of the relationships. In the next portion, we would like to focus on doing regressions, to analyze the variables of age vs depression, economic hardship vs depression, and age vs economic hardship and other variable pairs that can fit into regression modles.

### **Source**

We found the descriptions of datasets in R packages in<https://vincentarelbundock.github.io/Rdatasets/datasets.html>. By looking at the number of columns and rows, we found this Job2 data. Then, we installed the [mediation] package in R server.

##### **Appendix**

##### **Table 1: Statistics summaries of depression scores**

## min Q1 median Q3 max mean sd n missing

## 1 1.36 1.83 2.36 3 1.870033 0.5659093 899 0

##### **Figure 1: Histogram of Depression Status Scores**

##### **Table 2: Statistics Summaries of Economic Hardship**

## min Q1 median Q3 max mean sd n missing

## 1 2.33 3 3.67 5 3.024138 0.9864363 899 0

##### **Table 3: Statistics Summaries of Age**

## min Q1 median Q3 max mean sd n missing

## 17.48767 29.3 36.63562 44.62329 72.47945 37.56506 10.45143 899 0

##### **Figure 2: Histogram of Age**

##### **Table 4: Table of Gender**

##

## male female

## 417 482

##

## male female

## 0.4638487 0.5361513

##### **Table 5: Table of Marital Status**

##

## divrcd married nevmarr separtd widowed

## 163 408 279 30 19

##

## divrcd married nevmarr separtd widowed

## 0.18131257 0.45383760 0.31034483 0.03337041 0.02113459

##### **Table 6: Table of Race**

##

## non.white1 white0

## 152 747

##

## non.white1 white0

## 0.1690768 0.8309232

##### **Table 7: Table of Education**

##

## bach gradwk highsc lt-hs somcol

## 146 112 272 50 319

##

## bach gradwk highsc lt-hs somcol

## 0.16240267 0.12458287 0.30255840 0.05561735 0.35483871

##### **Table 8: Table of Income**

##

## 15t24k 25t39k 40t49k 50k+ lt15k

## 206 218 110 201 164

##

## 15t24k 25t39k 40t49k 50k+ lt15k

## 0.2291435 0.2424917 0.1223582 0.2235818 0.1824249

##### **Figure 3: Boxplot of Depression Status and Gender**

##### **Figure 4: Boxplot of Depression Status and Race**

##### **Figure 5: Boxplot of Depression Status and Income and ANOVA**

## Call:

## aov(formula = depress1 ~ income, data = jobsnew)

##

## Terms:

## income Residuals

## Sum of Squares 5.04777 282.53973

## Deg. of Freedom 4 894

##

## Residual standard error: 0.5621743

## Estimated effects may be unbalanced

## Df Sum Sq Mean Sq F value Pr(>F)

## income 4 5.05 1.262 3.993 0.00323 \*\*

## Residuals 894 282.54 0.316

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Tukey multiple comparisons of means

## 95% family-wise confidence level

##

## Fit: aov(formula = depress1 ~ income, data = jobsnew)

##

## $income

## diff lwr upr p adj

## 25t39k-15t24k -0.06293623 -0.212245823 0.0863733642 0.7785193

## 40t49k-15t24k -0.02472021 -0.206179931 0.1567395046 0.9959079

## 50k+-15t24k -0.20691107 -0.359257471 -0.0545646629 0.0020303

## lt15k-15t24k -0.05258940 -0.213398995 0.1082202038 0.8991964

## 40t49k-25t39k 0.03821602 -0.141496769 0.2179288016 0.9778336

## 50k+-25t39k -0.14397484 -0.294236219 0.0062865442 0.0677809

## lt15k-25t39k 0.01034683 -0.148488878 0.1691825455 0.9997751

## 50k+-40t49k -0.18219085 -0.364434529 0.0000528214 0.0501081

## lt15k-40t49k -0.02786918 -0.217244624 0.1615062584 0.9944868

## lt15k-50k+ 0.15432167 -0.007372037 0.3160153800 0.0696048

##### **Figure 6: Boxplot of Depression Status and Education and ANOVA**

## Call:

## aov(formula = depress1 ~ educ, data = jobsnew)

##

## Terms:

## educ Residuals

## Sum of Squares 1.52637 286.06113

## Deg. of Freedom 4 894

##

## Residual standard error: 0.5656668

## Estimated effects may be unbalanced

## Df Sum Sq Mean Sq F value Pr(>F)

## educ 4 1.53 0.3816 1.193 0.313

## Residuals 894 286.06 0.3200

##### **Figure 7: Boxplot of Depression Status and Marital Status and ANOVA**

## Call:

## aov(formula = depress1 ~ marital, data = jobsnew)

##

## Terms:

## marital Residuals

## Sum of Squares 2.61495 284.97255

## Deg. of Freedom 4 894

##

## Residual standard error: 0.5645894

## Estimated effects may be unbalanced

## Df Sum Sq Mean Sq F value Pr(>F)

## marital 4 2.61 0.6537 2.051 0.0854 .

## Residuals 894 284.97 0.3188

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### **Figure 8: Scatter plot of Depression Status and Age**

##### **Figure 9: Boxplot of Economic Hardship and Depression Status and ANOVA**

## Df Sum Sq Mean Sq F value Pr(>F)

## econ\_hard 1 29.66 29.657 103.1 <2e-16 \*\*\*

## Residuals 897 257.93 0.288

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### **Table 9: Summary Statistics of Depression Status and Occupations**

## occp min Q1 median Q3 max mean sd

## 1 clerical/kindred 1 1.360 1.820 2.27 3.00 1.823180 0.5409307

## 2 craftsmen/foremen/kindred 1 1.360 1.820 2.36 3.00 1.877938 0.6000763

## 3 laborers/service wrks 1 1.525 1.955 2.36 3.00 1.936429 0.5454691

## 4 manegerial 1 1.360 1.865 2.36 3.00 1.879107 0.5642311

## 5 operatives/kindred wrks 1 1.270 1.910 2.45 3.00 1.865806 0.6221078

## 6 professionals 1 1.450 1.910 2.27 3.00 1.900343 0.5551492

## 7 sales workers 1 1.400 1.730 2.36 2.91 1.829846 0.5808117

## n missing

## 1 217 0

## 2 97 0

## 3 84 0

## 4 168 0

## 5 93 0

## 6 175 0

## 7 65 0

##### **Figure 10: Boxplot of Marital and economic hardship and ANOVA**

## Df Sum Sq Mean Sq F value Pr(>F)

## marital 4 30.5 7.621 8.079 2.12e-06 \*\*\*

## Residuals 894 843.3 0.943

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##### **Table 10: Summary Statistics of Depression Status and Economic Hardship**

## econ\_hard min Q1 median Q3 max mean sd n missing

## 1 1 1.00 1.09 1.360 1.730 2.73 1.506316 0.5417278 19 0

## 2 1.330000043 1.00 1.09 1.450 2.000 2.91 1.609535 0.5689504 43 0

## 3 1.669999957 1.00 1.18 1.450 1.775 2.73 1.564000 0.5010862 55 0

## 4 2 1.00 1.18 1.550 1.910 2.82 1.630833 0.5127274 72 0

## 5 2.329999924 1.00 1.27 1.685 2.120 3.00 1.747750 0.5400961 80 0

## 6 2.670000076 1.00 1.36 1.820 2.270 2.82 1.816250 0.5403655 112 0

## 7 3 1.00 1.36 1.910 2.270 3.00 1.859764 0.5381397 127 0

## 8 3.329999924 1.00 1.50 1.910 2.360 3.00 1.941009 0.5415409 109 0

## 9 3.670000076 1.00 1.55 2.045 2.450 2.91 1.995341 0.5617009 88 0

## 10 4 1.00 1.64 2.180 2.640 3.00 2.134714 0.5662430 70 0

## 11 4.329999924 1.09 1.73 2.270 2.550 3.00 2.143000 0.5458760 50 0

## 12 4.670000076 1.20 1.90 2.090 2.450 2.82 2.115405 0.4325608 37 0

## 13 5 1.00 1.82 2.180 2.550 2.91 2.120270 0.5430750 37 0

##### **Figure 11 : Boxplot of Occupation and Economic Hardship**

##### **Table 11: Summary Statistics of Economic Hardship and Occupation and ANOVA**

## occp min Q1 median Q3 max mean sd

## 1 clerical/kindred 1.00 2.00 3.00 3.67 5 2.887742 1.0443565

## 2 craftsmen/foremen/kindred 1.00 2.67 3.33 3.67 5 3.154948 0.9852772

## 3 laborers/service wrks 1.33 2.33 3.00 3.67 5 3.087262 0.9662228

## 4 manegerial 1.00 2.33 3.00 4.00 5 3.061607 0.9800919

## 5 operatives/kindred wrks 1.00 2.33 3.00 3.67 5 3.053656 1.0174657

## 6 professionals 1.00 2.67 3.00 3.67 5 3.040171 0.8911761

## 7 sales workers 1.33 2.33 3.00 3.67 5 3.020462 1.0254061

## n missing

## 1 217 0

## 2 97 0

## 3 84 0

## 4 168 0

## 5 93 0

## 6 175 0

## 7 65 0

## Df Sum Sq Mean Sq F value Pr(>F)

## occp 6 6.4 1.0657 1.096 0.363

## Residuals 892 867.4 0.9724

##### **Figure 12: Boxplot of Race and Economic Hardship and T-test**

##

## Welch Two Sample t-test

##

## data: econ\_hard by nonwhite

## t = 3.4422, df = 205.07, p-value = 0.0006992

## alternative hypothesis: true difference in means is not equal to 0

## 95 percent confidence interval:

## 0.1363921 0.5021024

## sample estimates:

## mean in group non.white1 mean in group white0

## 3.289408 2.970161

##### **Table 12: Summary Statistics of Economic Hardship and Race**

## nonwhite min Q1 median Q3 max mean sd n missing

## 1 non.white1 1 2.67 3.33 4.00 5 3.289408 1.0576779 152 0

## 2 white0 1 2.33 3.00 3.67 5 2.970161 0.9631351 747 0