SAS® GLOBAL FORUM 2021

Paper #1197-2021

Pandemic Pandemonium

Hannah Flynt, Maryam Taherirani, Sean Everett, Trinh Phan;

Oklahoma State University

ABSTRACT

Losing a job is not only devastating to the individual and affected household, but is also a major economic issue. Okun's law, named after Yale Economist Arthur Okun, states that even a 1 percent increase in US unemployment causes a reduction in US GDP by 2 percent (Kenton, 2020). With the recent global health crisis, job loss has drastically increased and the goal of this paper is to explore publicly available demographic and infection rate data, inorder to understand the variables that most impact job loss and to predict the likelihood of losing one's job. Whether the household has experienced job loss is the target variable.

The final model, built in SAS Enterprise Miner, was a logistic regression model with a misclassification rate of 35%. Some identified variables of importance were Education, Income, Age, Health Insurance Coverage, Metropolitan Statistical Area, Government Response Index, Infection Rate and Number of Adults in the Household. The created model can be used as a resource for predicting geographic regions that are likely to experience increased levels in job loss based off the demographics of that region.

INTRODUCTION

COVID-19 is the worst global pandemic in generations, rivaling the Spanish flu of 1918- 1920. It has wrought death and economic destruction on a global scale not seen since WorldWar II. Communities nationwide implemented strict lockdown measures in an effort to curb the spread of the virus and to prepare the nation's healthcare system for the expected deluge of COVID infected patients. Schools shuttered and daycares closed, forcing millionsof parents to take extended leave from work, quit their jobs, or juggle the dueling responsibilities of distance learning and work. As a result, in the US, over 20 million people lost their jobs in the month of April, 2020 (Bureau of Labor Statistics, 2020), the single, largest monthly unemployment freefall in US history. The US unemployment rate skyrocketed from 3.5% in February to 14.7% in April, 2020 (Bureau of Labor Statistics, 2020).

PROBLEM

The unemployment rate affects many levels of society, from the individual to businesses, and, in the end, the national economy. When an individual loses their job, they then have less disposable income. Less disposable or personal income ultimately reduces economic growth and output (Hamel, 2020). Businesses also suffer as unemployment benefits are largely financed by taxes on businesses. This analysis aims to explore which factors have the highest probabilistic impact on whether an individual loses their job during the COVID-19 pandemic.

DATA

The primary dataset is from the Household Pulse Survey[1], which was created by the US Census Bureau, in collaboration with other federal agencies. The Household Pulse survey collects information from respondents about their pandemic experience in relation to employment, food scarcity, mental health, housing, health care coverage, and educational disruptions. The survey also includes geographic region and several demographic variables such as year born, marital status and gender. Several supplemental datasets were added to provide additional insights: the US Population Estimates[2] from the US Census Bureau, Job Industry[3] data from the US Bureau of Labor Statistics, COVID-19[4] cases from the NY Times COVID database, and US government policy and containment indices[5] from Oxford University's COVID Policy Tracker.

METHODS

DATA COLLECTION

This analysis focuses on Phase 1 of the survey results, which is over a 12-week time span that ranges from April 23 to July 21, 2020. The 12 weekly CSV files were merged together using a DATA MERGE step to create the primary SAS dataset. This resulted in a dataset with 93 variables and 1,088,314 records. In an effort to reduce the dataset to a manageable size, yet also be representative of regions commonly reported by federal agencies, the focuswas narrowed to metropolitan statistical areas (MSA) which reduced the dataset to 333k records.

From the NY Times COVID dataset, a derived variable was created: infection rate per 100,000 persons. Then from the US Census Bureau, the February 2020 employment totals by industry was used to create derived per 100,000-person employment variables. The finalsupplemental dataset incorporated was the Oxford University COVID policy dataset. This dataset provides policy rating indices at the state level.

After merging the aforementioned supplemental datasets by matching key fields, theresulting dataset had 108 variables and 333k records.

DATA CLEANING AND VALIDATION

The data was then cleaned to prepare for modeling. Since the goal of this project is predicting households impacted by work loss, all records were removed where the survey respondent answered "Retired" or "Choose not to work" as the reason for not working.

Next, the household work loss target variable, wrkloss, was transformed into a binary variable, where 0 implies "No, the household has not been impacted by work loss" and 1 implies "Yes, the household has been impacted by work loss". Then, all observations without a work loss response were eliminated.

For missing values, any variable with 50% or more missing responses were removed and the SURVEYIMPUTE PROC was used to replace the missing values using hotdeck imputation.

Variables that were not relevant to predicting work loss or variables that might bias the prediction, were removed. This included questions that asked whether people had money for food, or whether there was worry about losing their job.

[1] https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html

[2] https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/

[3] https://www.bls.gov/oes/

[4] https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

[5] https://raw.githubusercontent.com/OxCGRT/covid-policy-tracker/master/data/OxCGRT_latest.csv

After data cleaning process and variable reduction, the final dataset contains 49 variables and 273,984 observations.

For demographic analysis, four graphs were created to analyze job loss of participants by gender, age, education and income.

A key takeaway from the data is more females experience job loss than males. Among all ages surveyed, a large proportion of job loss occurs to persons above 36 years old. When analyzing job loss by education, the Graduates degreed group showed that only 36.5% of this category lost their job, while 66.9% of High School dropouts experienced job loss. As displayed by Figure 4, higher income workers retain their job more so than lower income workers, particularly compared to people whose income is less than $50,000. ANOVA and t-tests in Table 1, 2, 3, and 4 in the Appendix indicate statistically significant differences in job loss among the various groups analyzed.

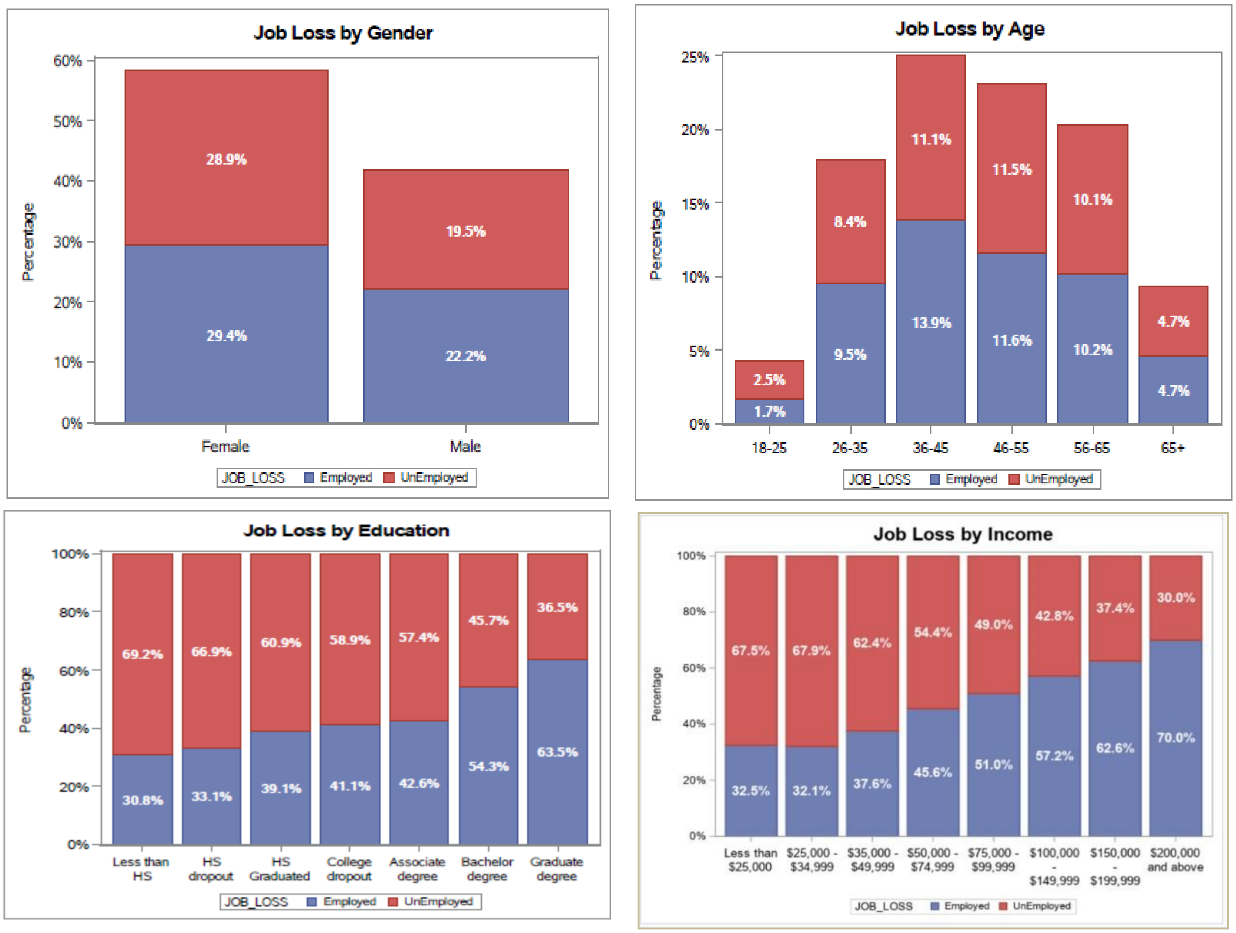


Figure 1, 2, 3 and 4. Job Loss by Gender, Age, Education and Income

ANALYSIS

The first step in data modeling was to reduce dimensionality and keep the most informative variables by using the following nodes in SAS Enterprise Miner: Partial Least Squares Regression (PLS), Least Angle Regression (LARS), Variable Clustering, and Least Absolute Shrinkage and Selection Operator (LASSO). Then the resulting selected data was partitioned to 70% training and 30% validation. Next, multiple models were trained, including Neural Network, Ensemble, Decision Tree, Gradient Boosting and Logistic Regression.

Comparison of the models' results shows that Variable Clustering, using four clusters, was the best for selecting relevant input variables. After running the model, the Logistic Regression model was selected as the champion model in terms of accuracy and interpretability. The misclassification rate of the champion model was 35% on the validation dataset, with sensitivity of 59% and specificity of 71.7%.

RESULTS

Some of resulting variables of importance at a significance level of 0.05 are summarized below.

- **Age**: Odds ratio of age is 1.003 which means that by being older by one year, the probability of losing a job will increase 1.003 times more.
- **Education**: this variable is categorical and the control specimen is the group with the highest degree (graduate degree). All odds ratios of the levels of this group are more than 1, meaning that the less educated, the more likely to lose their job.
- **Health Ins**: the odds ratio for the group of people who have health insurance coverage is 0.45 compared to the group of people who do not have this insurance.
- **Income:** this variable represents different ranges of income. It is concluded that higher income workers have less probability of losing a job.
- **Interest:** the odds ratio of this variable demonstrates that people who are more frequently interested in things around them, are less likely to lose their jobs compared to the ones who have the feeling of little interest in things.
- **Race:** There are four groups of races in the data; White, Black, Asian, and Others. The result shows that the Other group is more likely to lose the jobs; this result isconvincing since Race has a significant association with both Education and Income.
- **COVID19 infection rate**: although this is a significant variable, its odds ratio is close to 1, meaning that the infection rate of COVID 19 does not change the probability of losing a job.
- **Government-Response-Index:** the larger the index is in value, the more responsive the government is (more support from the government). Although the odds ratio is close to 1, it indicates that having a supportive government leads to less job loss.
- **Metropolitan Statistical Area:** this variable considers the effect of the region. The highest odds ratio in this group is for Detroit-Warren-Dearborn which is nearly 2. The mentioned result and the selection of MSA variable as a significant variable imply that influential factors leading to job loss may be different in different regions.

Although the Household Pulse survey was created in response to the COVID pandemic, similar survey datasets, such as the American Community Survey, exist on a national level which provide very similar yearly household demographic information. Consequently, the analysis could be applied at a non-COVID level, as well as expanded to include all regions ofthe US, and the results would be generalizable to represent job loss and the US population on a non-COVID level.

SUGGESTIONS FOR FUTURE STUDIES

Other results from this data suggest that Health Status and Martial Status effect on job loss. It shows that the people who are single or have good health are more likely to lose job. It is not intuitively clear why this would be the case and warrants further research. Moreover, it would be beneficial for future research to analyze job loss at a detailed occupation level and to consider communities at a more granular level. This would be useful for local governing bodies to have insight into providing mental health support, food resources, and benefits to the households likely to experience the loss of a job. In addition, further research should investigate the long-term effect on health and economics based off the indexes or levels of the government response for containment and information policies implemented on a local and national scale. This could help determine future response in the event of a global health emergency.

CONCLUSION

Based on research, almost 50% of households report that they have experienced some type of job loss during the COVID pandemic. Employers could consider providing employees with support for caregivers, either in the form of extended time off or subsidized daycare. Lower income and less educated individuals also report higher levels of job loss and this demographic is most reliant on recurring paychecks. Subsidy or stimulus payments to households making less than the Federal Poverty Level could help mitigate the long-term effect of unemployment on the economy, as well as provide immediate relief for individuals struggling for basic food security.

REFERENCES

BLS. (2020, 12 04). *Bureau of Labor Statistics.* Retrieved from Bureau of Labor Statistics: https://www.bls.gov/ces/publications/highlights/2020/current-employment-statistics-highlights-10-2020.pdf

Bureau of Labor Statistics. (2020, December 04). *TED: The Economics Daily.* Retrieved from Bureau of Labor Statistics: https://www.bls.gov/opub/ted/2020/unemployment-rate-rises-to-record-high-14-point-7-percent-in-april-2020.htm

Hamel, G. (2020, 12 1). *Unemployment and Fiscal Policy*. Retrieved from Chron: https://smallbusiness.chron.com/unemployment-fiscal-policy-12614.html

Kenton, W. (2020, December 2). *Okun's Law*. Retrieved from Investopedia: https://www.investopedia.com/terms/o/okunslaw.asp

Pew Research. (2020, 12 3). *Pew Research Center Social & Demographic Trends.* Retrieved from Pew Research Center: pewsocialtrends.org

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

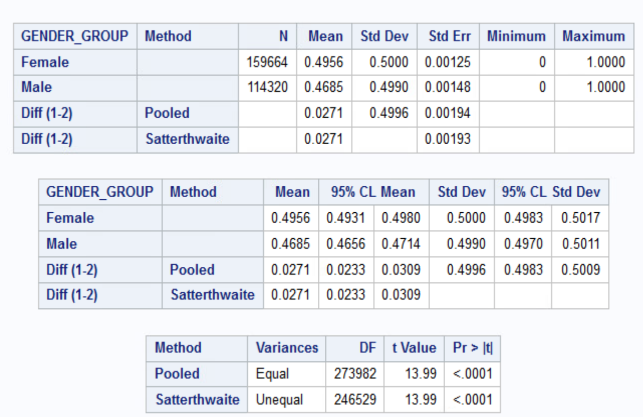
Hannah Flynt
Hannah.flynt@okstate.edu

Sean Everett
Sean.everett@okstate.edu

Maryam Taherirani
Maryam.taherirani@okstate.edu

Trinh Phan
Trinh.phan@okstate.edu

APPENDIX :

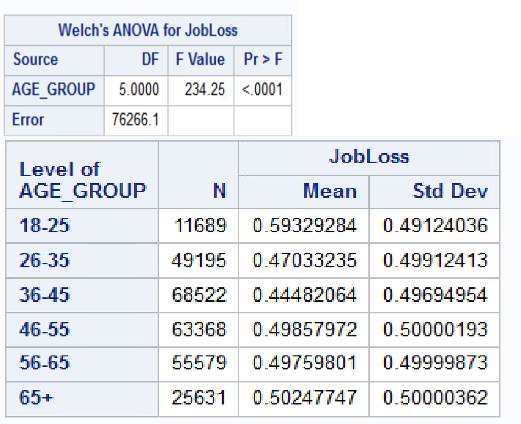


GENDER_GROUP	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female		159664	0.4956	0.5000	0.00125	0	1.0000
Male		114320	0.4685	0.4990	0.00148	0	1.0000
Diff (1-2)	Pooled		0.0271	0.4996	0.00194		
Diff (1-2)	Satterthwaite		0.0271		0.00193		

GENDER_GROUP	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		0.4956	0.4931	0.4980	0.5000	0.4983	0.5017
Male		0.4685	0.4656	0.4714	0.4990	0.4970	0.5011
Diff (1-2)	Pooled	0.0271	0.0233	0.0309	0.4996	0.4983	0.5009
Diff (1-2)	Satterthwaite	0.0271	0.0233	0.0309			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	273982	13.99	<.0001
Satterthwaite	Unequal	246529	13.99	<.0001

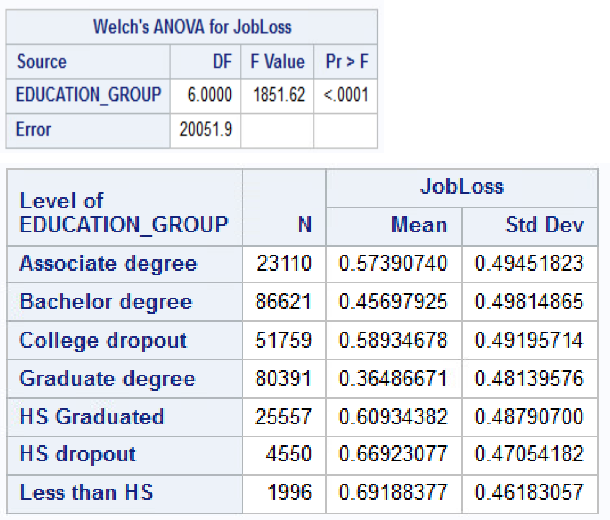
Table 1. T-test for Gender



Welch's ANOVA for JobLoss			
Source	DF	F Value	Pr > F
AGE_GROUP	5.0000	234.25	<.0001
Error	76266.1		

Level of AGE_GROUP	N	JobLoss	
		Mean	Std Dev
18-25	11689	0.59329284	0.49124036
26-35	49195	0.47033235	0.49912413
36-45	68522	0.44482064	0.49694954
46-55	63368	0.49857972	0.50000193
56-65	55579	0.49759801	0.49999873
65+	25631	0.50247747	0.50000362

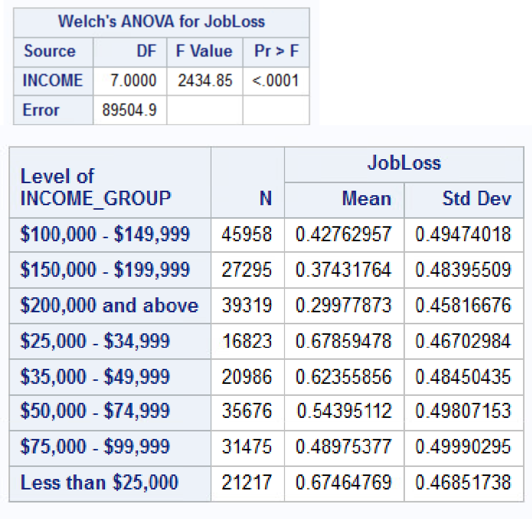
Table 2. ANOVA test for Age group



Welch's ANOVA for JobLoss			
Source	DF	F Value	Pr > F
EDUCATION_GROUP	6.0000	1851.62	<.0001
Error	20051.9		

Level of EDUCATION_GROUP	N	JobLoss	
		Mean	Std Dev
Associate degree	23110	0.57390740	0.49451823
Bachelor degree	86621	0.45697925	0.49814865
College dropout	51759	0.58934678	0.49195714
Graduate degree	80391	0.36486671	0.48139576
HS Graduated	25557	0.60934382	0.48790700
HS dropout	4550	0.66923077	0.47054182
Less than HS	1996	0.69188377	0.46183057

Table 3. ANOVA test for Education



Welch's ANOVA for JobLoss			
Source	DF	F Value	Pr > F
INCOME	7.0000	2434.85	<.0001
Error	89504.9		

Level of INCOME_GROUP	N	JobLoss	
		Mean	Std Dev
$100,000 - $149,999	45958	0.42762957	0.49474018
$150,000 - $199,999	27295	0.37431764	0.48395509
$200,000 and above	39319	0.29977873	0.45816676
$25,000 - $34,999	16823	0.67859478	0.46702984
$35,000 - $49,999	20986	0.62355856	0.48450435
$50,000 - $74,999	35676	0.54395112	0.49807153
$75,000 - $99,999	31475	0.48975377	0.49990295
Less than $25,000	21217	0.67464769	0.46851738

Table 4: ANOVA test for INCOME

Predicted Job Loss	Actual Job Loss		
	No (0)	Yes (1)	Total
No (0)	30,410	16,448	46,858
Yes (1)	11,982	23,358	35,340
Total	42,392	39,806	82,198

Table 5: Confusion Matrix of Validation set for Job Loss prediction

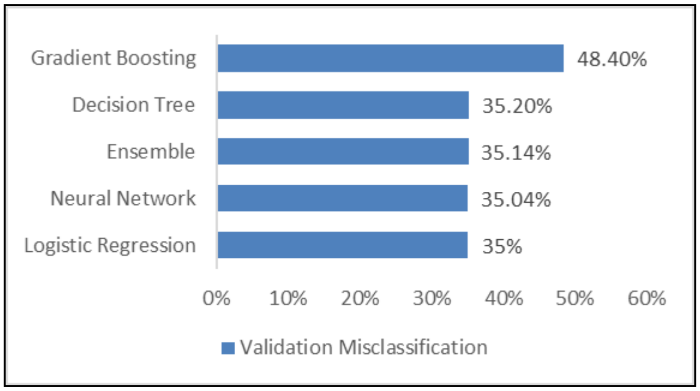
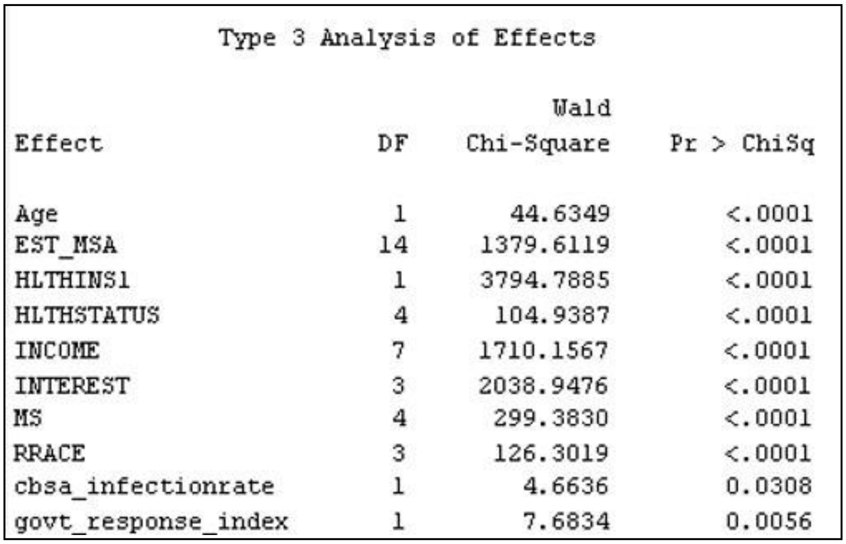


Figure 5. Model Comparison Validation Misclassification Rate

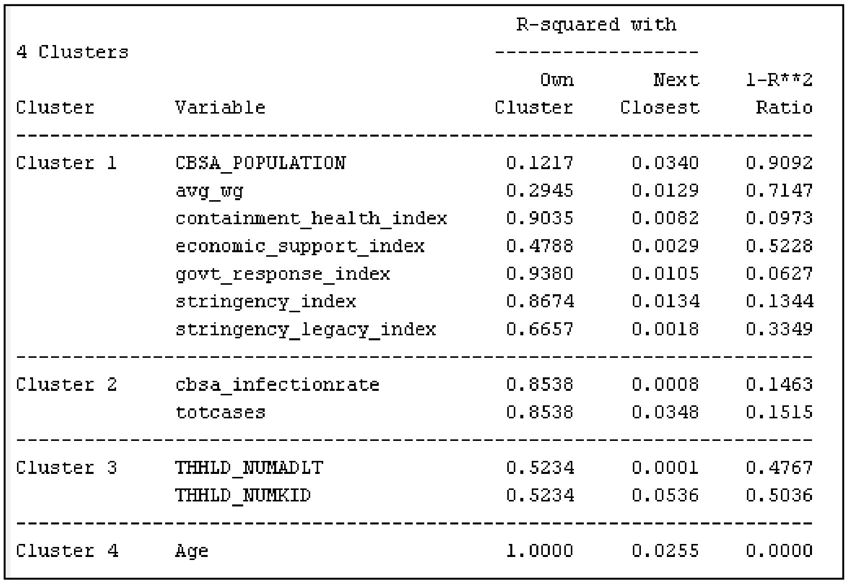


```
                    Type 3 Analysis of Effects

                                               Wald
Effect                          DF       Chi-Square      Pr > ChiSq

Age                              1          44.6349          <.0001
EST_MSA                         14        1379.6119          <.0001
HLTHINS1                         1        3794.7885          <.0001
HLTHSTATUS                       4         104.9387          <.0001
INCOME                           7        1710.1567          <.0001
INTEREST                         3        2038.9476          <.0001
MS                               4         299.3830          <.0001
RRACE                            3         126.3019          <.0001
cbsa_infectionrate               1           4.6636          0.0308
govt_response_index              1           7.6834          0.0056
```

Figure 6. Variable Significance



```
                                             R-squared with
4 Clusters                                ------------------
                                            Own        Next     1-R**2
Cluster        Variable                   Cluster    Closest     Ratio
------------------------------------------------------------------------
Cluster 1      CBSA_POPULATION             0.1217     0.0340    0.9092
               avg_wg                      0.2945     0.0129    0.7147
               containment_health_index    0.9035     0.0082    0.0973
               economic_support_index      0.4788     0.0029    0.5228
               govt_response_index         0.9380     0.0105    0.0627
               stringency_index            0.8674     0.0134    0.1344
               stringency_legacy_index     0.6657     0.0018    0.3349
------------------------------------------------------------------------
Cluster 2      cbsa_infectionrate          0.8538     0.0008    0.1463
               totcases                    0.8538     0.0348    0.1515
------------------------------------------------------------------------
Cluster 3      THHLD_NUMADLT               0.5234     0.0001    0.4767
               THHLD_NUMKID                0.5234     0.0536    0.5036
------------------------------------------------------------------------
Cluster 4      Age                         1.0000     0.0255    0.0000
```

Figure 7. Variable Clustering Result