HS256F - Healthcare Data Analytics and Data Mining

Professor Moaven Razavi

February 9, 2022

Homework 3

ANALYTICS OF PATIENTS AND CONSUMERS SURVEY

Group 4

Christianah Adeoya

Cyndi Ng

Siddhartha Kumar

Zilin Luo

INTRODUCTION

One way to thoroughly assess the quality of any system is to engage the users of the system. In the United States, the Centers for Medicaid and Medicare Services (CMS) sponsors one of such assessment tools called the Medicare Current Beneficiary Survey (MCBS) (CMS.gov, 2021). Having been in existence for over 30 years, the MCBS was designed as a "continuous, in-person, longitudinal survey of a representative national sample of the Medicare population" (CMS.gov, 2021).

A leading source of information pertaining to the administration, monitoring, and evaluation of Medicare, the MCBS according to the CMS (2021) determines for all Medicare beneficiaries, the expenditures and payment sources for all services provided, services not covered by the program. The survey also brings to bear the different types of insurance coverage of the Medicare beneficiaries, and how they relate to the sources of payment, as well as beneficiary satisfaction amongst other things (CMS.gov, 2021).

This report aims to analyze the MCBS Public-Use File collected for 2016 and assess it for certain relationships. A mix of statistical packages was used to carry out the analyses in this report including but not limited to STATA/SE 17.0 and Python. The data set had 12, 852 observations and it contained responses from those aged 65 years and above as well as other Medicare beneficiaries such as those with disabilities and some others. Figure 1 shows the distribution of the survey's participants across these different groups.

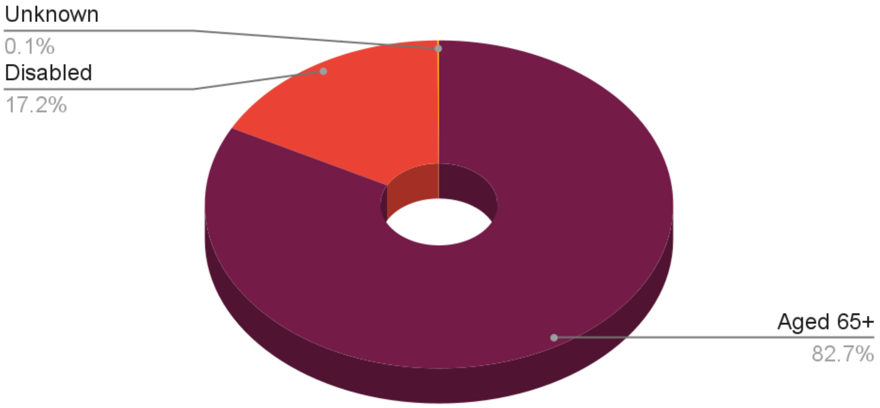


Figure i.1: Chart showing the distribution of the survey participants

PART 1 - Racial Disparity In Ability To Pay For Care

This portion of the report seeks to analyze the data to see if there exists any form of racial disparity in the number of beneficiaries who were unable to access care in the last year due to cost, particularly the non-Hispanic white and non-Hispanic black racial groups. This analysis was carried out only on the portion of the data representing beneficiaries over the age of 65, using Fisher's exact test.

Method: Using the STATA/SE 17.0 package, the data for the respondents over 65 years was filtered out of the main data set and saved as a subset. Using the subset, all entries other than non-Hispanic white and non-Hispanic black from the variable representing race were dropped, and all other variables apart from 'Yes' and 'No' in the variable representing delay in care due to cost in the last year were also dropped. The Fisher's exact test was then run for the delay in care due to cost against the race variable in a 2 x 2 table.

Results and Analysis: Using the subset with respondents aged 65 years and over, 5.96%(539) of the 10,629 individuals experienced a delay in care due to cost in the past year. In assessing the racial distribution in the group, the non-Hispanic whites made up 77.19%(8,205) of those aged 65 and over.

Figure 1.1: Visual representation of patients over 65 years who experienced delay in care due to cost

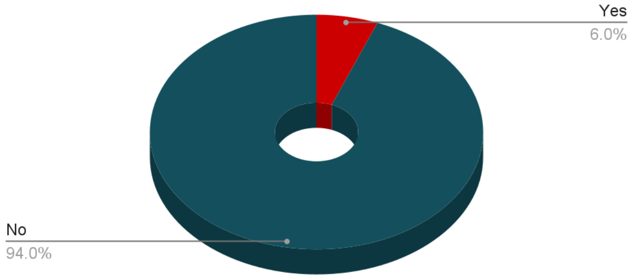
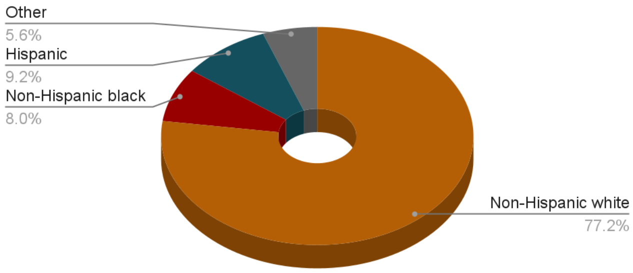


Figure 1.2: Visual representation of the racial distribution in patients over 65 years



Delay in care due to cost in the last year	Non-Hispanic White	Non-Hispanic Black
Yes	462	77
No	7,731	772

Table 1: Delay in care due to cost in the last year by race (particularly Non-Hispanic white and Non-Hispanic black)

The statistical analysis showed that there was a highly significant difference based on race in the delay of care due to cost in the last month ($p \leq 0.001$), focusing only on non-Hispanic whites and non-Hispanic blacks with the former experiencing more delays than the former.

PART 2 - Gender Differential In Healthcare Utilization

Detailed herein is an analysis to understand the rate of healthcare utilization and how it varies based on gender.

Method: Using the STATA/SE 17.0 package, the data for the respondents over 65 years was filtered out of the main data set and saved as a subset. Using the subset, a proxy variable was created to represent health utilization as a count variable. This was then used to calculate the weighted average for each gender after cross-tabulating the gender variable with the total number of hospital visits in the current year.

Results and Analysis: Table 2 below shows the details of the analysis. A total of 4,659 males and 5,970 females were taken into account for this analysis. The data showed males having a weighted average of 4.92 while females had a weighted average of 5.26.

Male	Health Utilization	Total Office visits in 2016	Frequency	Weighted average
4,659	0	2,116	0	
	3	962	2,886	
	8	746	5,968	
	13	429	5,577	
	18	216	3,888	
	23	200	4,600	
			22,919	**4.92**
Female	0	2,508	0	
5,970	3	1,258	3,774	
	8	1,044	8,352	
	13	604	7,852	
	18	277	4,986	
	23	279	6,417	
			31,381	**5.26**

Table 2: Weighted average of frequency of hospital visits per gender

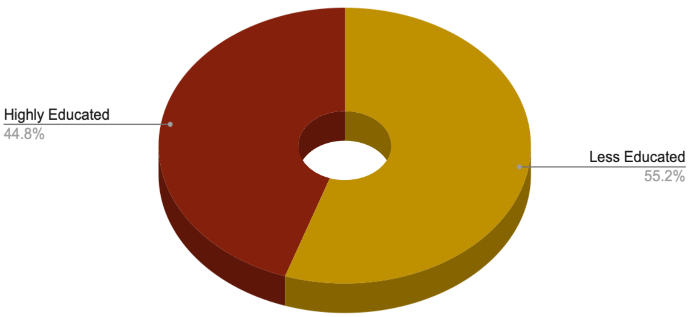
PART 3 - Relationship Between Education And Health

This section of the report sought to explore the relationship between a consumer's level of education and their health status, particularly as it related to obesity and it's more dangerous counterpart, extreme/high-risk obesity. This analysis, premised on the theory by Folland et al. (2017) that education increases the chances of an individual living better and making better lifestyle choices, seeks to explore if indeed this data confirms the theory.

Method: Using the STATA/SE 17.0 package, the data for the respondents over 65 years was filtered out of the main data set and saved as a subset. Using the subset, a dummy variable was created with two categories to represent only the healthy respondents and those with BMI values from 30Kg/m^2 and above. A dummy variable was also created to recategorize the respondents' level of education, grouping those with values less than 3 as 'Less Educated' and those with values of 3 or more as 'Highly Educated'. A Fisher's exact test was run with level of education against BMI to obtain a 2 x 2 table.

Results and Analysis: From the subset created, respondents with BMI values classified as 'Obese' made up 25.72% (3,192) of the total distribution aged 65 and over, while 4.89% (607) were classified as being at extreme/high-risk obesity. In assessing the level of education, 55.20% (7,063) were less educated while 44.80% (5,733) were highly educated.

Figure 4: Visual representation of the education level of the respondents



Level of Education	Healthy	Obesity
Less Educated	4,599	2,208
Highly Educated	4,014	1,591

Table 3: Delay in care due to cost in the last year by race (particularly Non-Hispanic white and Non-Hisp

The Fisher's test carried out showed that there was a highly significant difference in the obesity levels between the highly educated and the less educated, with the lesser-educated respondents having higher instances of obesity and high-risk obesity than their highly educated counterparts ($p \leq 0.001$).

PART 4 - Relationship Between Obesity And Anxiety/Depression

This portion of the report seeks to explore if any, the link between obesity and anxiety/depression. As some pieces of literature have often linked obesity to certain other health conditions, this analysis will explore using the data set provided the existence of such links.

Method: Using the data set provided, a new variable is created to represent those with BMI values of 4 and 5 representing Obesity and Extreme/High-Risk Obesity respectively from the original BMI variable. A proxy variable was also created to represent depression. Two groups were created in this variable by joining categories 1, 2, and 3 as the healthy group and joining categories 4 and 5 as the obese group suffering from depression.

Results and Analysis: By deploying Fisher's Exact test (p<0.001), we find that there is high significance between obesity and the incidence of anxiety/depression among all age groups.

	Obesity	Non-obesity
Anxiety/Depression	1,366	2,039
Non-anxiety/non-depression	2,446	6,593

Table 4: Cross-tabulation of depression against obesity

PART 5 - Gender Differential In The Reciprocal Relationship Between Obesity And Depression

This part of our analysis focuses on how the obesity-depression causality varies with gender factored into this equation. We try to determine if this incidence is higher in females than males, or vice-versa.

Method: Using STATA, we cleaned the data by categorizing BMI into healthy and obese. Those categorized under 1, 2, and 3 were healthy while 4 and 5 were in the obese category. Furthermore, a Fisher's exact test was conducted to figure out if gender shows any significance to the obesity-depression causality.

Results and Analysis: By deploying Fisher's Exact test (p<0.001), we find that there is high significance between BMI and the incidence of anxiety/depression. However, this incidence is seen higher in females than in males.

Males	Low BMI	High BMI
Anxiety/Depression	779	491
Non-anxiety/non-depression	3236	1189

Females	Low BMI	High BMI
Anxiety/Depression	1260	875
Non-anxiety/non-depression	3357	1257

Table 5: Depression vs BMI per gender

PART 6 - Loneliness and health

This portion of the analysis sought to explore the relationship between loneliness, measured as living alone with no family, and the respondent's perception of their health relative to others their age.

Method: First, the data were filtered to ensure only individuals aged 65 and over were being studied. Then data wrangling was carried out to remove rows with null or non-numeric values in the variables of choice - General health condition compared to others their age and Marital status. Next, these variables were recoded to Health Condition and Loneliness for use in Fisher's exact 2 x 2 test.

Result and Analysis: The cross-tabulation of Health Condition and Loneliness is shown below:

Loneliness	**Good Health**	**Poor/Fair Health**
Living Alone	4,033	1,049
With Family	4,590	895

Table 6: Cross-tabulation of health condition by loneliness

By employing the Fisher_exact function, the p-value was obtained which showed a highly significant difference in the health condition of those who live alone as against those who live with their families ($p \leq 0.001$).

PART 7 - Loneliness And The Risk Of Depression

This section details the exploration of the risk of the seniors being diagnosed with depression in different living conditions, specifically living alone, without a family or any effective social relationships, and living with a family.

Method: Using Python, the data for individuals with age 65 and over was filtered out to create a sub-set of the data. In this sub-set, all rows with null values and meaningless values (D, R) were dropped. Selected variables relevant to this analysis - Marital Status and History of ever having depression were assessed and then used.

Result and Analysis: Using Fisher's exact test, a highly significant difference was established in having depression between those who live with their families and those who live alone ($p \leq 0.001$).

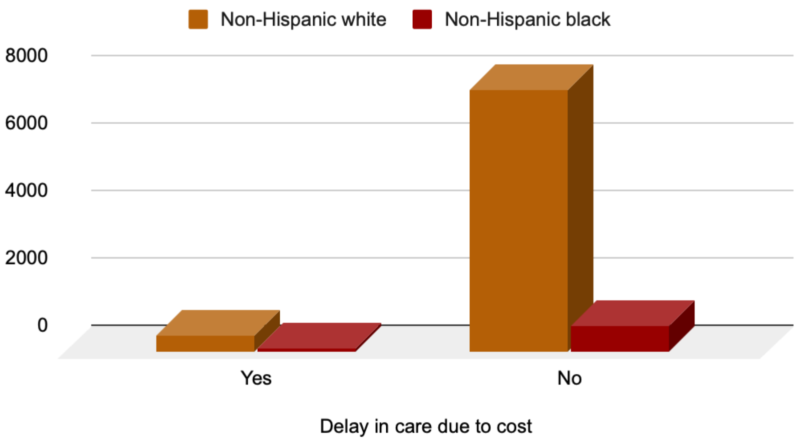
	Live with Families	Live Alone
Have Depression	992	1,255
No Depression	4,513	3,839

Table 7: Cross-tabulation of living status with depression

DISCUSSION AND LIMITATIONS

Representing 82.7% of the main data set, some of the analyses in this report was carried out using a subset containing those aged 65 years and over. It can be established from the analysis carried out that there were more non-Hispanic white respondents who experience delays in accessing care due to their inability to pay for it than the non-Hispanic black respondents.

Graph 1: Delay in care due to cost vs race

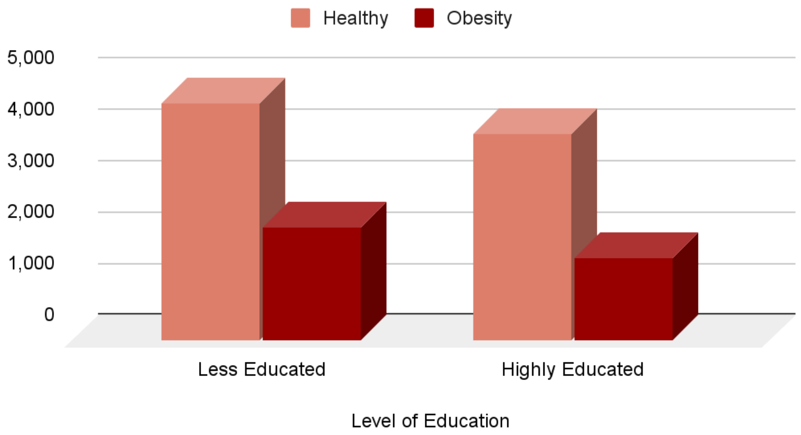


From both Table 1 and Graph 1, it is observed that only a small percentage of the respondents experienced this delay however, in that small number, the non-Hispanics were significantly more than the non-Hispanic blacks.

The analysis to determine what gender, if any, has a higher health-seeking behavior through analyzing hospital visits proved that females, in fact, visit the hospital a lot more frequently than the males with weighted average visits of 5.26 and 4.92 respectively. This can be attributed to the fact that genetically, women have more reasons to visit the hospital, stemming from incidences related to reproductive issues, child-bearing, and issues arising from their hormones. There also lies herein the fact that women most naturally tend to want to handle issues arising with their health almost immediately, without any cajoling than men do.

Similar to the case of the racial disparity, when the level of education variable was run against the BMI of the respondents aged 65 years and over, there was a highly statistically significant difference in the incidence of obesity based on the level of education. This confirms the theory by Folland et. al in their book on 'The Economics of Health and Health Care' (2017) where they pointed out that people with higher levels of education tend to make better lifestyle choices and as such have lesser

Graph 3: Relationship between level of education and obesity



incidences of obesity than those with lesser education.

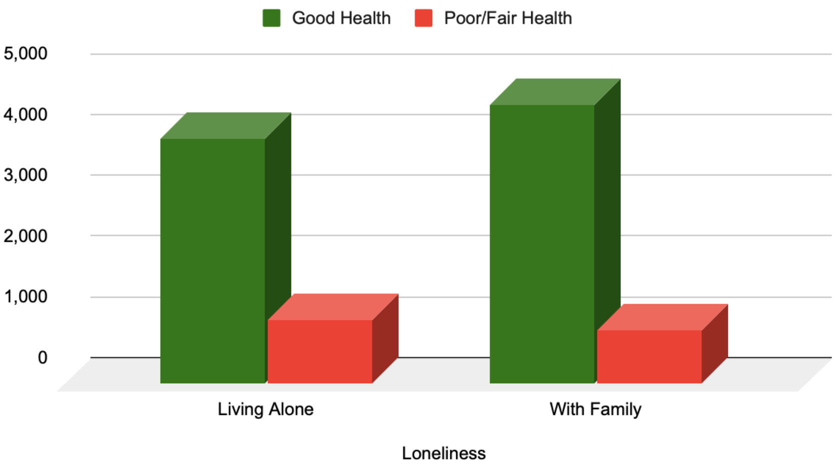
The analysis clearly shows that there is high significance between the incidence of depression in obese people. This also means that the null hypothesis can be clearly rejected in this scenario.

Furthermore, the factoring in of gender in this causality has thrown up some interesting observations. While showing the high significance and with the null hypothesis rejected, it is clear that obesity in females (40.99%) leads to higher anxiety than in males (38.66%). The anxiety could be due to the physical appearance created because of obesity.

In the analysis on loneliness, there was a highly significant difference between the health condition of those who live alone and those who live with family, with the latter having the better health condition

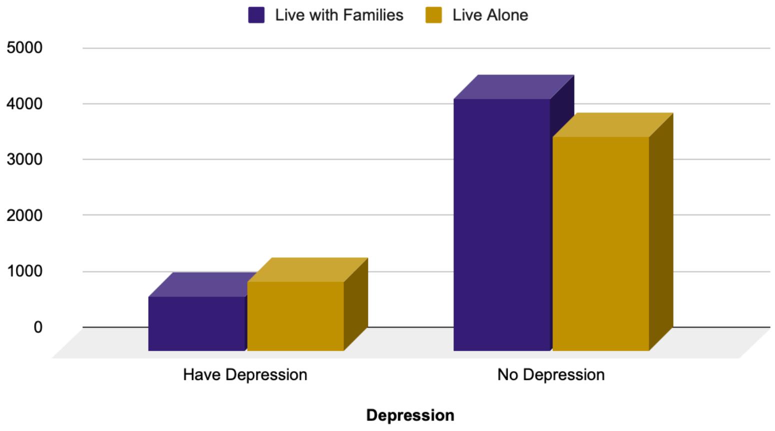
With the result obtained in the analysis, the null hypothesis can be rejected, concluding that living with family/a partner will lead to better health conditions.

Graph 6: Relationship between health condition and loneliness



Further analysis into loneliness and depression showed a clear significant difference between those who live alone and those who live with their families with regard to depression. In light of this, the null hypothesis that the risk of depression is independent of living with family can be rejected. The data shows sufficient evidence that whether elderly people live alone or with a family does in fact affect the risk of them coming down with depression. As seen in Graph 7, a greater proportion of people who live by themselves have been diagnosed with depression, so it can be concluded that the

Graph 7: Relationship between loneliness and depression



seniors who live alone are more likely than their peers who live with families to suffer from depression.

CONCLUSION

In conclusion, this report detailed the various patterns observed from the MCBS 2016 data, identifying relationships and trends between certain variables which are essential in not just understanding the current status of the Medicare beneficiaries but more importantly in making decisions and shaping policies that directly make for better health outcomes and quality of life for Medicare beneficiaries.

REFERENCES

CMS.gov. (2021, December 1). *MCBS Public Use File*. CMS. Retrieved February 18, 2022, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/MCBS-Public-Use-File

Folland, S., Goodman, A. C., & Stano, M. (2017). *The Economics of Health and Health Care: International Student Edition, 8th Edition* (8th ed.). Taylor & Francis Group.

APPENDIX 1

```
clear

capture log close

**Set directory
cd "/Users/bloomn/Library/Mobile Documents/com~apple~CloudDocs/Personal Documents/Heller GHPM/Classes & Reading Resources/Spring '22/HS 256f - Healthcare Data Analytics & Data Mining/Class 4/MCBSpuf2016"

**Create log
log using "PS_HW3.log", replace

**Import dataset
import excel "/Users/bloomn/Library/Mobile Documents/com~apple~CloudDocs/Personal Documents/Heller GHPM/Classes & Reading Resources/Spring '22/HS 256f - Healthcare Data Analytics & Data Mining/Class 4/MCBSpuf2016/PUF2016.xlsx", sheet("Sheet1") firstrow

**FIlter for age>65
keep if ADM_H_MEDSTA ==1

su

**Save subset
save "puf16_medsta.dta", replace

**Check the DEM_RACE variable and drop the unwanted values (3 and 4)
tab DEM_RACE

drop if DEM_RACE==3
drop if DEM_RACE==4
tab DEM_RACE

**Check the ACC_HCDELAY variable and drop the unwanted values (D and R)
tab ACC_HCDELAY

drop if ACC_HCDELAY== "D"
drop if ACC_HCDELAY== "R"

tab ACC_HCDELAY

**Run the Fisher's Exact 2 x 2 Test
tab ACC_HCDELAY DEM_RACE, exact
```

APPENDIX 2

```
clear

**Load the data set
use puf16_medsta

**Create and recode proxy variable for utilization - ADM_UTIZ**
generate ADM_UTIZ=.
replace ADM_UTIZ=0 if ADM_H_PHYEVT==0
replace ADM_UTIZ=3 if ADM_H_PHYEVT==1
replace ADM_UTIZ=8 if ADM_H_PHYEVT==2
replace ADM_UTIZ=13 if ADM_H_PHYEVT==3
replace ADM_UTIZ=18 if ADM_H_PHYEVT==4
replace ADM_UTIZ=23 if ADM_H_PHYEVT==5

**View Gender variable
tab DEM_SEX

**Cross-tabulate to get the gender distribution for ADM_H_PHYEVT
tab DEM_SEX ADM_H_PHYEVT
```

>>Transferred and completed in Excel<<

APPENDIX 3

```
clear

 **Load the original dataset
use PUF2016

**See BMI Variable - HLT_BMI_CAT's freq
tab HLT_BMI_CAT

**Create new variable (BMI_CAT_OB) with values of 4 and 5 for HLT_BMI_CAT to use in relationship
test
generate BMI_CAT_OB=.
replace BMI_CAT_OB=0 if HLT_BMI_CAT==4
replace BMI_CAT_OB=1 if HLT_BMI_CAT==5

tab BMI_CAT_OB

**See education variable - DEM_EDU, drop observations D, N and R and convert variable to numeric
tab DEM_EDU
drop if DEM_EDU=="D" | DEM_EDU=="N" |DEM_EDU=="R"
destring DEM_EDU, replace

**Create new variable (DEM_EDU_RECAT) from DEM_EDU to use in Fisher’s test
generate DEM_EDU_RECAT=.
replace DEM_EDU_RECAT=0 if DEM_EDU <3
replace DEM_EDU_RECAT=1 if DEM_EDU >=3

label define rect 0 "Less Educated" 1 "Highly Educated"
label values DEM_EDU_RECAT rect

tab DEM_EDU_RECAT

**Run the Fisher's Exact 2 x 2 Test
tab DEM_EDU_RECAT BMI_CAT_OB, exact
```

APPENDIX 4

```
generate BMI_45=.
replace BMI_45=4 if hlt_bmi_cat ==4
replace BMI_45=5 if hlt_bmi_cat ==5

gen depress=.
encode hlt_ocdeprss , generate(DEPRESSION)
replace depress=1 if DEPRESSION ==1
replace depress=2 if DEPRESSION ==2

tabulate depress BMI_45, exact row
```

APPENDIX 5

```
tabulate depress BMI_45 if dem_sex==1, exact row

tabulate depress BMI_45 if dem_sex==2, exact row
```

APPENDIX 6

```
import pandas as pd
import numpy as np

raw_df = pd.read_csv('/Users/cynding/Desktop/Brandeis 2022 Spring/HS 256F - Healthcare Data
Analytics/HW3/MCBSpuf2016/puf2016.csv')

# Filter 65+ agegroup
survey_df = raw_df[raw_df['ADM_H_MEDSTA']==1].reset_index()
survey_df.shape
print(survey_df['HLT_GENHELTH'].unique())
print(survey_df['DEM_MARSTA'].unique())
survey_df.dropna(subset = ['HLT_GENHELTH','DEM_MARSTA'], inplace=True)
survey_df.shape
survey_df = survey_df[(survey_df['HLT_GENHELTH'] != 'D') &
                (survey_df['HLT_GENHELTH'] != 'R') &
                (survey_df['DEM_MARSTA'] != 'D') &
```

```
              (survey_df['DEM_MARSTA'] != 'R')]

# Recode HLT_GENHELTH to HealthCondition
survey_df['HLT_GENHELTH'] = survey_df['HLT_GENHELTH'].astype(int)
survey_df['HealthCondition'] = np.where(survey_df['HLT_GENHELTH'] >= 4, 'PoorFairHealth',
'GoodHealth')

# Recode DEM_MARSTA to Loneliness
survey_df['DEM_MARSTA'] = survey_df['DEM_MARSTA'].astype(int)
survey_df['Loneliness'] = np.where(survey_df['DEM_MARSTA'] == 1, 'WithFamily', 'LivingAlone')

crosstab1 = pd.crosstab(survey_df['Loneliness'],survey_df['HealthCondition'])
print(crosstab1)
print(crosstab1.iloc[0,0],crosstab1.iloc[0,1],crosstab1.iloc[1,0],crosstab1.iloc[1,1])
from scipy.stats import fisher_exact
oddsratio,pvalue=fisher_exact([[crosstab1.iloc[0,0],crosstab1.iloc[0,1]],
[crosstab1.iloc[1,0],crosstab1.iloc[1,1]]])
pvalue
```

APPENDIX 7

```
df_i =
pd.read_csv("puf2016.csv",usecols=["ADM_H_MEDSTA","HLT_GENHELTH","DEM_MARSTA","HLT_OCDEPRSS"])
df = df_i[df_i["ADM_H_MEDSTA"]==1]
df.shape
df.dropna(subset=['DEM_MARSTA','HLT_OCDEPRSS'],inplace = True)
drop_value = ["D","R"]
df.drop(df[df['DEM_MARSTA'].isin(drop_value)].index, inplace = True)
df.drop(df[df['HLT_OCDEPRSS'].isin(drop_value)].index, inplace = True)
df.shape
df["DEM_MARSTA"] = df["DEM_MARSTA"].astype(int)
df["HLT_OCDEPRSS"] = df["HLT_OCDEPRSS"].astype(int)
WF_D = (df["DEM_MARSTA"]==1) & (df["HLT_OCDEPRSS"]==1)
LA_D = (df["DEM_MARSTA"]!=1) & (df["HLT_OCDEPRSS"]==1)
WF_ND = (df["DEM_MARSTA"]==1) & (df["HLT_OCDEPRSS"]!=1)
LA_ND = (df["DEM_MARSTA"]!=1) & (df["HLT_OCDEPRSS"]!=1)
gr = np.array([[WF_D.sum(),LA_D.sum()],[WF_ND.sum(),LA_ND.sum()]])
df= pd.DataFrame(gr, columns=['With Family','Live Alone'])
df.index = ['Depression','No Depression']
df
oddscratio, pvalue =stats.fisher_exact(df)
pvalue
```