**PQHS 515: Health Care Analytics**

**Fall 2023**

**Homework 2**

**Due Date: September 29**

*Please submit via Canvas, attaching your syntax (.sas, .log, .lst files) will give the opportunity for additional credit.*

For this homework assignment you will be using death certificate data from the years 1998 and 1999. These modified datasets have 60 variables about Ohio persons who died in those years. The 1998 dataset utilizes ICD-9 codes for cause of death, which consist of a 3 – 5 digit number. Beginning in 1999, Ohio switched to ICD-10 codes for cause of death. ICD-10 codes begin with a letter.

**Death certificate files are located in: /meta/db4/epbi515/homework/hmwk2/data**

**Answer the following questions:**

1. Explore the data. Run proc contents, proc print a sample, and run some proc freq statements to get a feel for the data. Run any additional exploratory procedures you feel like. (No answer required for this one.)
2. Identify the 5 leading causes of death for all persons in Ohio in **1998**, using the variable “cause” which is a ICD-9 code representing the underlying cause of death. Data will often be presented this way and require you to use a codebook or external documentation. Resources for mortality codes can be found online. (5 pts)

|  |  |  |
| --- | --- | --- |
| ICD-9 code | Cause of Death (description) | # of deaths |
| 410 | Acute myocardial infarction | 9,819 |
| 1629 | Malignant neoplasm of bronchus and lung, unspecified | 7,623 |
| 4140 | Coronary atherosclerosis of unspecified type of vessel, native, or graft | 6,479 |
| 4149 | Chronic ischemic heart disease, unspecified | 5,514 |
| 436 | Acute, but ill defined, cerebrovascular disease | 4,794 |

1. When using administrative data over time, it’s important to realize that coding can change. In 1999, Ohio switched to ICD-10 codes to identify cause of death. Identify the 5 leading causes of death for all persons in Ohio in **1999**, using the variable “cause” which is a ICD-10 code representing the underlying cause of death. (5 pts)

|  |  |  |
| --- | --- | --- |
| ICD-10 code | Cause of Death (description) | # of deaths |
| I251 | Atherosclerotic heart disease of native coronary artery | 11,835 |
| I219 | Acute myocardial infarction, unspecified | 9,740 |
| C349 | Malignant neoplasm of unspecified part of bronchus or lung | 7,597 |
| I64 | Stroke, but not specified as hemorrhage or infarction | 5,253 |
| J449 | Chronic obstructive pulmonary disease | 4,667 |

1. Identify the 5 leading causes of death for all persons in Ohio in **1999**, using the NCHS 113 Selected Causes of Death classification. The variable is “nchs\_1” and the data is presented as a numerical value. It groups ICD-10 codes into more meaningful categories. Resources to identify cause of death for NCHS 113 classification can be found online. (5 pts)

|  |  |  |
| --- | --- | --- |
| NCHS 113 # | Cause of Death | # of deaths |
| 55 | All other forms of chronic ischemic heart disease | 12,738 |
| 52 | Acute myocardial infarction | 9,746 |
| 61 | Cerebrovascular disease | 8.048 |
| 25 | Malignant neoplasms of trachea, bronchus, and lung | 7,633 |
| 95 | All other diseases (residual) | 7,141 |

1. Again, identify the 5 leading causes of death for persons in Ohio in **1999**, but this time use the NCHS 50 classification (called “Rankable Causes of Death”). The NCHS further classifies the 113 causes down to 50 – for instance it groups all cancers together. The variable in the dataset is “nchs\_2”. (5 pts)

|  |  |  |
| --- | --- | --- |
| NCHS 50 # | Cause of Death | # of deaths |
| 23 | Diseases of heart | 35,090 |
| 15 | Malignant neoplasms | 25,908 |
| 99 | Unknown | 10,992 |
| 25 | Cerebrovascular diseases | 8.048 |
| 30 | Chronic lower respiratory diseases | 5,815 |
| 18 | Diabetes mellitus | 3,851 |

Teaching Point (not a question): “What are the top 5 causes of death in Ohio?” seems like a straight-forward question but as you can see from questions #3, #4, and #5 we can come up with different answers depending on how we group and classify diseases. When working with large databases, you will often have to make choices about how to classify cases and controls or how to classify procedures. This should be done “a priori” to avoid accusations of data dredging (i.e. redefining your cohort until you get an answer that is what you are looking for).

1. Identify the 3 leading causes of death for each of the following age groups in **1999**, using the **NCHS 113** classification. (3pt)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 – 14 | 15 – 34 | 35 – 64 | 65 and older |
| Leading cause of death | Certain conditions originating in perinatal period | Motor vehicle crash | Malignant neoplasms of trachea, bronchus, and lung | All other forms of chronic ischemic heart disease |
| 2nd leading cause | Congenital malformations, deformations, and chromosomal abnormalities | Intentional self-harm (suicide) by discharge of firearms | Acute myocardial infarction | Acute myocardial infarction |
| 3rd leading cause | Symptoms, signs, and abnormal clinical and lab findings | Assault (homicide) by discharge of firearms | All other forms of chronic ischemic heart disease | Cerebrovascular disease |

1. Suppose you wanted to examine the relationship between marital status and dying of acute myocardial infarction. Your explanatory variable is marital status (in the dataset) and your outcome of interest is death due to MI (a binary variable: 0=alive, 1=dead). What are some problems with using death certificate data alone to answer this question? (5pt)  
     
   A problem that we would have is that death certificates only show marital status at time of death, not in the past where their status may have been different. Another problem is that we don’t know any other issues that the person may have had that may contribute to their death due to MI, such as having high blood pressure.
2. In death certificate data, the ICD9 code and ICD10 codes for injury and poisoning related deaths are actually represented by Ecodes (external cause of death). These include all ICD9 codes between 800-999, and all ICD10 codes that start with a V, W, X, or Y.

Find the number of deaths by poisonings and deaths by firearms in **1998 and 1999** (combined) and breakdown by intent. Display your data by age groups. Also determine what percentage of total deaths this represents for each age group. (6pt)

Poisonings

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Unintentional | | Homicide (intentional inflict on others) | | Suicide  (intentional self-inflicted) | |
| Age Group | # of deaths | % of all deaths for age grp | # of deaths | % of all deaths for age grp | # of deaths | % of all deaths for age grp |
| 1 – 14 | 3 | 0.29% | 1 | 0.10% | 0 | 0.00% |
| 15 – 34 | 179 | 3.39% | 0 | 0.00% | 94 | 1.78% |
| 35 – 64 | 538 | 1.31% | 2 | 0.005% | 240 | 0.58% |
| 65 | 28 | 0.02% | 2 | 0.0012% | 55 | 0.03% |

Firearm-related deaths

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Unintentional | | Homicide (intentional inflict on others) | | Suicide  (intentional self-inflicted) | |
| Age Group | # of deaths | % of all deaths for age grp | # of deaths | % of all deaths for age grp | # of deaths | % of all deaths for age grp |
| 1 – 14 | 8 | 0.76% | 14 | 1.34% | 12 | 1.15% |
| 15 – 34 | 17 | 0.32% | 364 | 6.88% | 380 | 7.19% |
| 35 – 64 | 13 | 0.03% | 201 | 0.49% | 597 | 1.45% |
| 65 | 4 | 0.002% | 21 | 0.012% | 370 | 0.22% |

1. As the previous example shows, one can determine the intent of injury and poisoning related deaths using ICD9 and ICD10 ecode variables. Suppose you are interested in studying utilization of mental health services by victims of suicide in the year prior to death. You want to compare across different mental illness diagnoses (depression, schizophrenia, etc.). Assuming you have access to other administrative data sources: (2 pages max) (6pt)

Describe your approach to the analysis.

What additional data source(s) might you use?

What are the strengths and weaknesses of these data sources?

Comment on the strengths and weaknesses of the approach you chose.

I would begin by getting patient IDs of people who were victims of suicide from the death certificate dataset as well as inpatient data, which would allow us to get data on peoples’ demographics and various diagnoses. Additionally, we can get HCUP data and link that with the IDs as well to get the victims’ past expenditures.

At this point the data would look like: patient ID – demographics – health diagnoses – past medical expenditures. Since we have already subsetted patient IDs to those who were victims of suicide, the first subset we have to do is to get past expenditures in the past year prior to death. Assuming that there is a date of expenditure variable within HCUP, we can find this by getting all expenditures between the date of death and a year prior. Since we’re only interested in utilization of mental health services, we can subset all expenditures to those related to mental health services. Next, we can subset the data again to patients who have mental illness diagnoses of interest. With all this subsetting, our data ow looks like: patient ID – demographics – mental illness diagnoses of interest – past year expenditures on mental health services.

We can now begin to analyze the data with all this information to investigate prior year utilization of mental health services by victims of suicide. Given that demographics typically has some sort of influence within health data, it’s likely that we will want to include the information within our analysis. There are various ways for us to analyze the data, with the simplest being an ANOVA test where we can compare the mental health service utilization for each group of mental illness diagnoses.

With these data sources, there are some strengths and weaknesses to consider. For death certificate data, there aren’t any weaknesses, with its strengths being that we know the date and cause of death. Typically, suicide deaths are easier to recognize, so we don’t really need to worry about a misdiagnosis. For the inpatient data, a strength is that we can get patients’ demographic information and various diagnoses. However, the diagnoses may also be a weakness as there’s always a chance of misdiagnosis, especially regarding mental health. Another issue is with how you quantify if the patient does or doesn’t have a diagnosis, as many diagnoses may be listed so you may only look at the first X diagnoses. The HCUP data is useful in that we can see a patient’s medical expenditures in the past year. A weakness may be how they categorize the expense as to whether or not it’s related to mental health services. Additionally, HCUP isn’t an exhaustive data set, so we may also be missing additional medical expenditures.

One of the strengths of an ANOVA test would be useful in this scenario in that we can compare the average medical expenditures of multiple groups of diagnoses. It is also fast and efficient to run, given that it isn’t very computationally intensive. A weakness is that while it can tell if there are statistically significant differences between groups, it doesn’t tell why. As such, post hoc analysis would need to be conducted in the event that there are differences between groups. Another weakness is that because the dataset will likely be very large, there’s a decent chance that each group is ‘statistically significant’ from each other because of the sample size. As such, confidence intervals may need to be calculated to actually be able to differentiate between the diagnosis groups.