BIOST 515 HW 6

Eliza Chai

03/02/2022

Responses

1. Visualisation of county-level mortality rates vary over time and associated with gender and age

Figure 1 shows the scatterplot panel of log number of deaths (for that county, sex, age, and year) versus log population size (stratified by age and sex in the county in the middle of the year) associated with gender and age groups. We can see a positive relationship between log number of deaths and log population size. For male and female subpopulations in the same age group, the trend between the number of deaths and population size appears to be similar (i.e. the slopes or mortality rates are similar among females and males of the same age group).

Figure 2 shows the scatterplot panel of the number of deaths (for hat county, sex, age, and year) versus year ranging from 2010 to 2019 by age groups. We observe that the over 85-years-old age group has the highest number of deaths in all years from 2010 to 2019. The 75-84 years-old age group has the second-highest number of deaths in all years from 2010 to 2019. The number of deaths in all years is concentrated below 1000 deaths for all age groups.

Figure 3 shows the lineplot panel of Gender and age-stratified mortality rates over time for Washington State counties. The Loess-smoothed mortality rates over all counties with confidence intervals are shown.

2. The fitted Poisson regression model is as follows:

$log(E(\hat{deaths}/popsize)) = 2.436 - 0.004 * year + 0.287 * 1_{gender} + 0.763 * 1_{age:65-74} + 1.770 * 1_{age:75-84} + 2.946 * 1_{age:over85}$.

$log(E(\hat{deaths}/popsize))$ is the log estimated county-level mortality rate. $year$ is the year ranging from 2010 to 2019. $1_{gender} = 1$ if the gender of a participant is male, $1_{gender} = 0$ if the gender of a participant is female. $1_{age:65-74} = 1$ if the age group of a participant is in 65-74 years-old group, $1_{age:65-74} = 0$ otherwise. $1_{age:75-84} = 1$ if the age group of a participant is in 75-84 years-old group, $1_{age:75-84} = 0$ otherwise. $1_{age:over85} = 1$ if the age group of a participant is the over 85-years-old group, $1_{age:over85} = 0$ otherwise

The log estimated mortality rate (number of deaths per population for that county, sex, age, and year) among a group of women in the age group of 55-64 years-old and 0 calendar year is 2.436. (We estimate that among the female population in the age group of 55-64 years-old in year 0, the observed county-level mortality rate is $e^{2.436} = 11.42$ deaths per population size). When comparing county-level mortality rates for the same gender and same age group but differing by 1 calendar year, we estimate that the mortality rate is $e^{-0.004} = 0.996$ times greater in the group with a higher calendar year. In addition, we estimated that the county-level mortality rate is $e^{0.287} = 1.332$ times greater in the male population compared to the female population when comparing mortality rates for the same age group and in the same year. The multiplicative difference in mortality rate comparing population in the age group of 65-74 years-old to a population in the 55-64 years-old group with the same gender and the same year is $e^{0.763} = 2.144$ times greater for the population in the age group of 65-74 years-old. When comparing the mortality rate for the population of the same gender

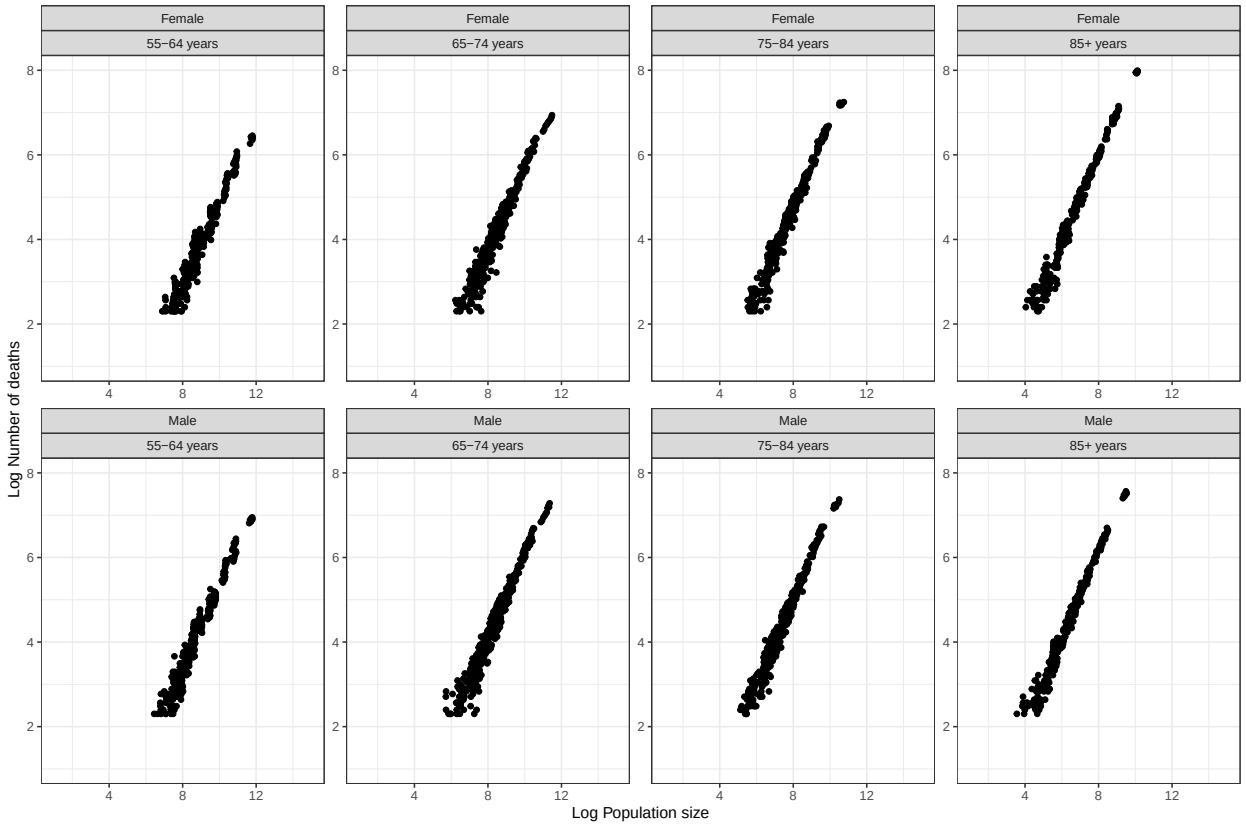

Figure 1: Scatterplot of Log county-level mortality associated with gender and age group in the WONDER dataset

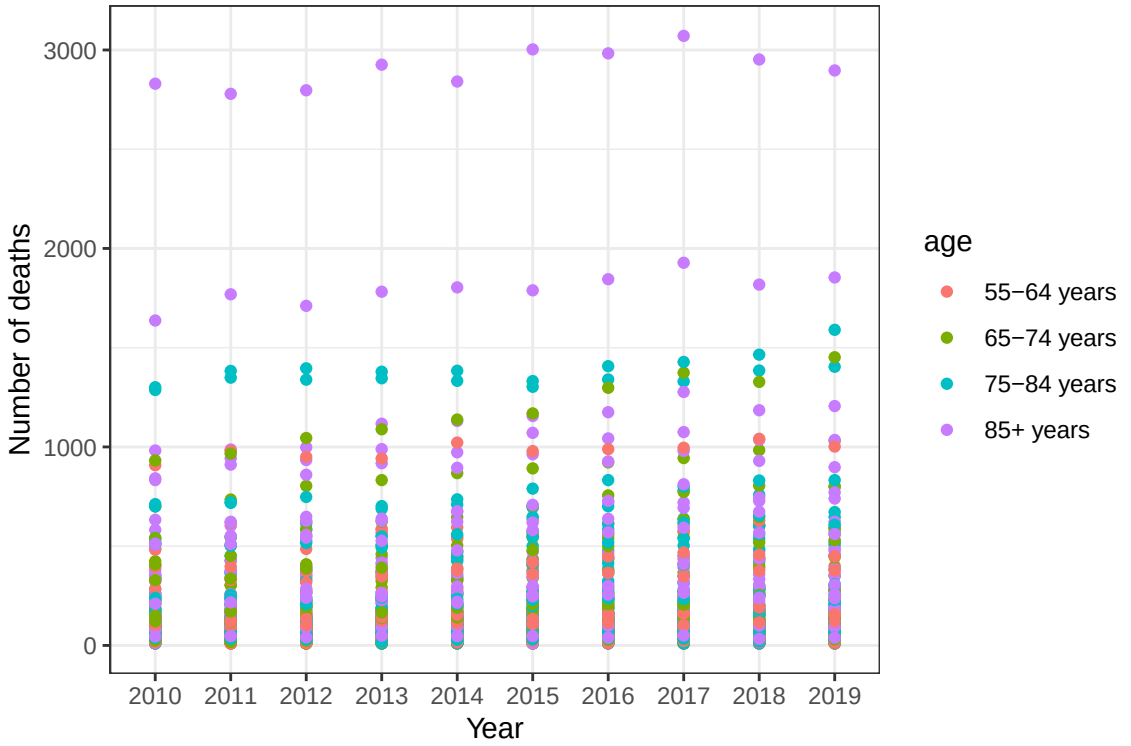

Figure 2: Scatterplot of county-level mortality over time by age in the WONDER dataset

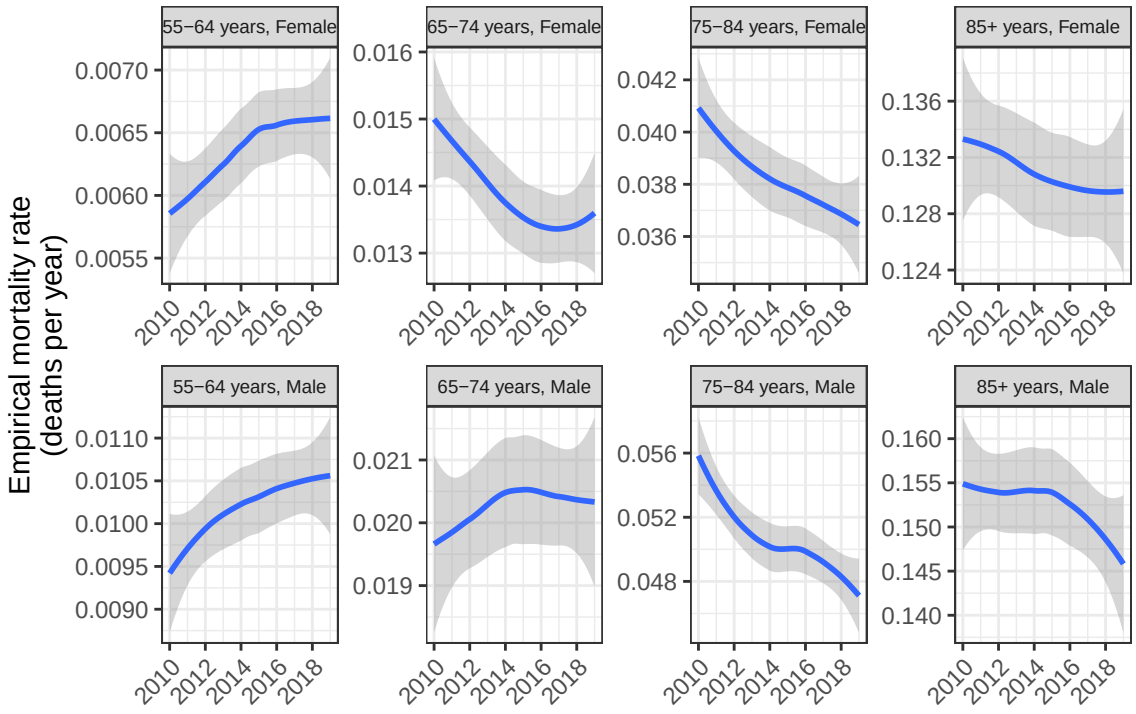

Figure 3: Gender and age-stratified mortality rates over time for Washington State counties. Loess-smoothed mortality rates over all counties are shown

and in the same calendar year, we estimate that the population in the age group of 75-84 years-old is $e^{1.770}$ = 5.870 times greater than the 55-64 years-old group. Lastly, when comparing the mortality rate for the population of the same gender and in the same calendar year, we estimate that the population in the age group over 85 years-old is $e^{2.946}$ = 19.02 times greater than the 55-64 years-old group.

3. We fit a Poisson regression model with log county-level mortality rate as the predictor and time (in the calendar year), gender, and age group as the response using robust-Wald test standard errors and performed tests at a 5% significance level. When comparing county-level mortality rate for the same gender and same age group but differing by one calendar year, we estimate that the mortality rate is $e^(- 0.004) = 0.996$ times greater in the group with a higher calendar year (95% CI with robust-Wald test: 0.993, 1.000 fold). We reject the null hypothesis of no association between county-level mortality rate and time (p = 0.0254) and conclude that there is evidence for an association between mortality rate and time. Thus, based on the model from (2), there is evidence that mortality rates are changing over time.

4. In treating the age groups as effect modifiers, we fit a model with an interaction between year and the indicators of three age groups. The newly fitted Poisson regression model is as follows:

$log(E(\hat{deaths}/popsize))$ = -14.691 - 0.005 $*$ $year$ + 0.287 $*$ 1_{gender} + 19.388 $*$ $1_{age:65-74}$ + 39.573 $*$ $1_{age:75-84}$ + 10.956 $*$ $1_{age:over85}$ - 0.009 $*$ $1_{age:65-74}$ $*$ $year$ - 0.019 $*$ $1_{age:75-84}$ $*$ $year$ - 0.004 $*$ $1_{age:over85}$ $*$ $year$

$log(E(\hat{deaths}/popsize))$ is the log estimated county-level mortality rate. $year$ is the year ranging from 2010 to 2019. $1_{gender} = 1$ if the gender of a participant is male, $1_{gender} = 0$ if the gender of a participant is female. $1_{age:65-74} = 1$ if the age group of a participant is in 65-74 years-old group, $1_{age:65-74} = 0$ otherwise. $1_{age:75-84}$ = 1 if the age group of a participant is in 75-84 years-old group, $1_{age:75-84} = 0$ otherwise. $1_{age:over85} = 1$ if the age group of a participant is the over 85-years-old group, $1_{age:over85} = 0$ otherwise.

We fit a Poisson regression model for log county-level mortality rate as the outcome, including gender, year, age groups, and three interaction terms between year and age groups variables as predictors. We used the robust-Wald test to perform inference on the parameters of our model to study the effect of different age groups on modification of the trends in mortality rates over time. We find that the difference between the multiplicative differences of all the different age groups was statistically significant (p = 0.0003). Therefore, we reject the null hypotheses that different age groups have the same effect on county-level mortality rate over time. We conclude that there is evidence of the effect modification of different age groups in the association between county-level mortality rates adjusted for gender over time.

5. From the data, not all county-gender-year-age tabulation groups are present. The lower number of deaths count reported is 10 deaths, therefore the threshold of suppressing counts is 10 deaths. County has 39 levels, gender has 2 levels, age has 4 levels, and the year has 10 levels. $39 * 2 * 4 * 10 = 3120$ Theoretically, there should be 3120 rows in the dataset, however, there are only 2709 rows in the WONDER dataset, meaning that 411 rows have been suppressed because of the low death counts.

6. From the dataset, we observe that the suppression of observations on death counts (lower than ten deaths) is more common in smaller or more rural counties. Counties such as Adams County, Wahkiakum County, and Sun Juan County are generally smaller in population size. Therefore, it is reasonable that these counties have lower death counts. However, the suppression of observations is on death counts, not in mortality rate, which is the estimated outcome of our models. As a result, the suppression of these observations or missing data from counties with potentially low mortality rates would lead to an overestimation in our models. We would be consistently overestimating the effect of county-specific mortality rate over time and the effect modification of age groups on the trend of mortality rate over time. On the other hand, these counties could have a high mortality rate as they have both low deaths counts and low population size. In this case, we would be consistently underestimating both the association of count-specific mortality rate over time and the effect of different age groups on the trend of mortality rate over time. Therefore, the suppression of observation could lead to a larger bias when estimating the parameters in our model in both part (2) and part (4).

7. It is important to assume observations are being independent to obtain reasonable estimates and valid statistical inference since many statistical tests rely on this assumption. We need consistency to show that an estimate using observation can approach the true value of the parameter. With independent observations, it is easier to guarantee consistency. For instance, if the county-level mortality rates in men and women are dependent, then the Poisson regression model we used in the above analysis would not be a valid model. However, in the WONDER dataset, we do have dependency among observations. The death counts for the same county over time (from 2010 and 2019) are correlated. The population in Adam County in 2010 and the population in Adam County in 2012 are likely to be the same. In addition, there could be correlations between the population of different age groups over time. For instance, in 2010, a 60 years-old individual in the 55 - 64 years-old group would be included in the 65 - 74 years-old group in 2015. Thus, there is a correlation among deaths counts of the different age groups over time. These correlated observations would impact our model and could lead to incorrect predictions if not adjusted accordingly in our analysis.

Code Appendix

```
### Setting up the packages, options we'll need:
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(rigr)
### ------------------------------------------------------------
### Reading in the data.
wonder <- read_csv("~/Desktop/UW MS Biostat/BIOST Winter 2022/Biost 515/R dataset/cdc-wonder-wa.csv")
### ------------------------------------------------------------
### Q1
### Visualisation of how county-level mortality rates vary over time and
### associated with gender and age

### Change type of gender and age to factor variables
wonder <- wonder %>%
        mutate(gender = as.factor(gender)) %>%
        mutate(age = as.factor(age))
### Scatterplot of number of deaths vs. population size by gender and age group
wonder %>%
        ggplot(aes(x = log(pop), y = log(deaths)))+
        geom_point() +
        facet_wrap(c("gender","age"), scales="free", nrow=2) +
        labs (y = "Log Number of deaths", x = "Log Population size") +
        xlim (1, 15) +
        ylim(1, 8) +
        theme_bw()
### Scatterplot of number of deaths vs. year ranging from 2010-2019 by age groups
wonder %>%
        ggplot(aes(x = as.factor(year), y = deaths, col=age), alpha = 0.5) +
        geom_point() +
        labs (y = "Number of deaths", x = "Year") +
        theme_bw()
### Line plot of mortality rate vs. year ranging from 2010-2019 by age and sex groups
wonder %>%
        mutate(rate = deaths/pop) %>%
        mutate(category = paste(age, gender, sep = ", ")) %>%
        group_by(county, gender, age) %>%
        ggplot(aes(x = year, y = rate)) +
        geom_smooth(span=1) +
        ylab("Empirical mortality rate\n(deaths per year)") +
        scale_x_continuous(breaks = c(2010, 2012, 2014, 2016, 2018, 2020)) +
        theme_bw() +
        xlab("") +
        theme(strip.text = element_text(size = 7)) +
        theme(axis.text.x=element_text(angle=45, hjust=1),
              legend.position="none") +
        facet_wrap(~category, scales="free", nrow=2, dir="v") +
        NULL
### ------------------------------------------------------------
### Q2
### Here's a Poisson regression model for wonder dataset using rigr:regress
```

```r
### mortality rate as response and time, gender, age group as predictors
mod1 <- regress("rate",
                deaths ~ age + gender + as.numeric(year),
                data = wonder,
                offset = log(pop))
mod1
### Alternative way to fit Poisson regression model for wonder dataset using glm
### mortality rate as response and time, gender, age group as predictors
mod2 <- glm(deaths ~ year + gender + age,
            family = poisson(link = "log"),
            data = wonder,
            offset = log(pop))
mod2 %>% summary %>% coef %>% round(3)
### ------------------------------------------------------------
### Q3
### Here's the coefficient of Poisson model for wonder dataset using rigr:regress
mod1 %>% coef %>% round(3)
### Alternative way: coefficients of Poisson model for wonder dataset using glm
mod2 %>% summary %>% coef %>% round(3)
mod2 %>%
        confint(parm = "year") %>%
        exp %>%  signif(3)
### ------------------------------------------------------------
### Q4
### Here's the coefficient of Poisson model for wonder dataset using rigr:regress
### mortality rate as response, time, gender as predictors, and age group as effect modifier
mod3 <- regress("rate",
                deaths ~ gender + as.numeric(year)*age,
                data = wonder,
                offset = log(pop))
mod3
### Alternative way to fit Poisson regression model for wonder dataset using glm
### mortality rate as response, time, gender as predictors, age group as effect modifier
mod4 <- glm(deaths ~ gender + age*year,
            family = poisson(link = "log"),
            data = wonder,
            offset = log(pop))
mod4 %>% summary %>% coef %>% round(3)
### Run ANOVA test with likelihood ratio test
anova(mod2, mod4, test = "LRT")

### ------------------------------------------------------------
### Q5
### Sort wonder dataset by deaths in ascending order
wonder %>% arrange(deaths)
### Looking at the structure of wonder dataset
str(wonder)
### ------------------------------------------------------------
```