Lab 2: Poisson Regression and Relative Risk

SOLUTIONS

24 September, 2021

Objectives

In this lab, we will be working on:

1. improving our familiarity with dplyr to manipulate data sets;
2. improving our familiarity with ggplot2 to visualize infectious disease data;
3. understanding and applying Poisson regression to infectious disease data in the US;
4. understanding and interpreting Poisson regression parameter estimates in the context of relative risk;
5. connecting our model results with the Poisson distribution.

Getting Started

Before beginning work on the lab, please clear your environment by clicking on the broom icon in the upper right window of your RStudio session. Also click Session > Restart R to start a new R session.

1. We will be working with some data, called `us_contagious_diseases` available through the *dslabs* package. This is not a package we have used before, so you will have to install it. Make sure you only install this once (so comment out the code line you used to install the package after it is installed.) You can write this line of code right after the line that says INSTALL dslabs package in the code chunk below.

```
str(us_contagious_diseases)
```

2. Now that we have loaded the data, we should find out what is stored in the `us_contagious_diseases` object. Use the `str` function to find out what kind of an object we have and how the information is stored in it.

```
## 'data.frame':    16065 obs. of  6 variables:
##  $ disease        : Factor w/ 7 levels "Hepatitis A",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ state          : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year           : num  1966 1967 1968 1969 1970 ...
##  $ weeks_reporting: num  50 49 52 49 51 51 45 45 45 46 ...
##  $ count          : num  321 291 314 380 413 378 342 467 244 286 ...
##  $ population     : num  3345787 3364130 3386068 3412450 3444165 ...
```

```
us_contagious_diseases %>%
  str
```

3. Last class, we used pipes (`%>%`) and `mutate()` to add a column to a data frame; pipes are part of the *dplyr* package. We can use pipes more generally, as in the code chunk below. Notice that this gives the same result as in your previous code chunk.

```
## 'data.frame':    16065 obs. of  6 variables:
##  $ disease        : Factor w/ 7 levels "Hepatitis A",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ state          : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year           : num  1966 1967 1968 1969 1970 ...
##  $ weeks_reporting: num  50 49 52 49 51 51 45 45 45 46 ...
##  $ count          : num  321 291 314 380 413 378 342 467 244 286 ...
##  $ population     : num  3345787 3364130 3386068 3412450 3444165 ...
```

```
us_contagious_diseases %>%
  pull(var=disease) %>%
  levels
```

4. The data set we have contains yearly incidence for seven different infectious diseases in the United States from 1966 to 2011. It would be nice to know what diseases are included. We can use pipes to do this, along with `pull` and `levels`.

```
## [1] "Hepatitis A" "Measles"     "Mumps"       "Pertussis"   "Polio"
## [6] "Rubella"     "Smallpox"
```

5. Comparing raw case counts between two states, for example, is not necessarily the best way to compare disease incidence. What should I account for to create some measure that is more comparable? What would this look like mathematically? Is this a rate, ratio, or proportion? Consider reporting this quantity per 10,000 people. This is a rate. We are reporting the new cases per year, and we are assuming that the population is reported as the average population in a year for each state.

```
us_contagious_diseases <- us_contagious_diseases %>%
  mutate(
    rate=count/population*10000
  )
```

6. Using pipes and the mutate function, let's create a new column in `us_contagious_diseases` that captures the quantity you described in Question 5. You should name it either rate, ratio, or proportion, depending on what quantity you identified. *This is a rate, assuming that the denominator is the average population in each state during the corresponding year.*

7. There are 52 weeks in a year. Do we have reporting for 52 weeks in every year and every state for every disease? If the answer is "no", what might this mean for case counts in terms of comparison year-to-year?. If you find an example helpful for building intuition, consider case counts in a year where we have reporting for 49 weeks versus one where we have reporting for 52 weeks. No, if we do not have 52 weeks of reporting, we would expect our case counts to be under-reported.

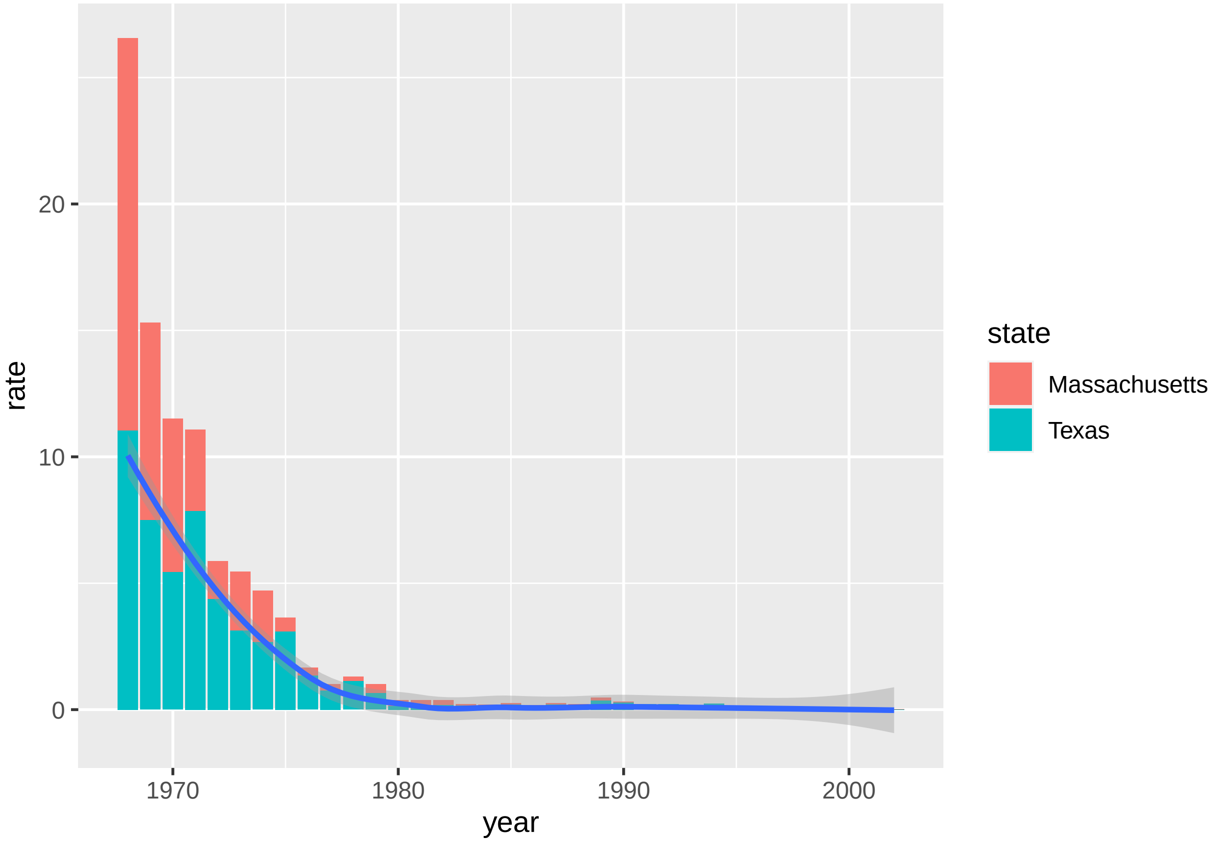
```
us_contagious_diseases <- us_contagious_diseases %>%
  mutate(
    rate=count/population*(52/weeks_reporting)*10000
  )
```

8. How would you adjust for the weeks of reporting? Modify your code in Question 6 to make this adjustment to the measure of frequency that you created.

```
# us_contagious_diseases %>%
#   filter(disease == "Mumps" & state == "Massachusetts" | state == "Texas") %>%
#   mutate(rate = count / population * 10000 * 52 / weeks_reporting) %>%
#   ggplot(aes(x=year, y=rate)) +
#   geom_col(aes(fill=state)) +
#   # geom_point() +
#   geom_smooth()

us_contagious_diseases %>%
  filter(disease == "Mumps" & (state == "Massachusetts" | state == "Texas")) %>%
  mutate(rate = count / population * 10000 * 52 / weeks_reporting) %>%
  ggplot(aes(x=year, y=rate)) +
  geom_col(aes(fill=state)) +
  # geom_point() +
  geom_smooth()
```

9. Here is some code that does some of things we have talked about, plus demonstrates the flexibility of piping. It might be helpful to use part of it in Question 10. Be careful, though - think carefully about what you might expect this code to include by way of states and diseases, and what it actually does.

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (position_stack).
```



```r
ma_tx_mumps <- us_contagious_diseases %>%
  filter(disease == "Mumps" & state == "Massachusetts" | disease == "Mumps" & state == "Texas") %>%
  mutate(rate = count / population * 10000 * 52 / weeks_reporting)
```

10. The data frame `us_contagious_diseases` contains information for 50 US states plus the District of Columbia. Create a new data frame with only information for Mumps cases in Texas and Massachusetts. Call it `ma_tx_mumps`. You should use pipes and the filter function. Consider looking at the help documentation for filter if needed.

Poisson (log-linear) model

The general form of the Poisson model is

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

where β_0 is the intercept, β_1 through β_p are the parameter coefficients that we need to estimate for the 1 through p explanatory variables. For today, we are only going to have only two explanatory variables ($p = 2$) - time and state.

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

When dealing with population/time adjusted rates, we need to add an offset term to the right side:

$$\log(\lambda) = \log\left(\frac{\text{population}}{\text{weeks reporting}/52}\right) + \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Model assumptions are:

1. **Poisson response.** The response variable is a count per unit time/space.
2. **Independence.** The observations are independent of one another.
3. **Mean=variance.** The mean of a Poisson random variable is equal to its variance.
4. **Linearity.** The log of the mean rate, $\log(\lambda_i)$ must be a linear function of the explanatory variable(s).

```
# test <- glm(count ~ year + state + offset(log(population*weeks_reporting/52)),
#             data=ma_tx_mumps,
#             family = poisson(link="log"))
# summary(test)
```

11. We can fit this model with the `glm()` function. This works much like the `lm` function, but we also have to specify a `family` argument, as well as adding an `offset()` term to the specification of the model formula. For help, look up the help documentation (`?glm`).

```
ma_tx_mumps_new <- ma_tx_mumps %>%
  filter(weeks_reporting != 0)

test <- glm(count ~ year + state + offset(log(population*weeks_reporting/52)),
            data=ma_tx_mumps_new,
            family = poisson(link="log"))
summary(test)
```

12. You should get an error when you try to fit this model as is. This is an error caused by something in the data. Think about the offset term - what variables are included? Inspect the range of each one to determine what the source of the error is.

```
##
## Call:
## glm(formula = count ~ year + state + offset(log(population *
##     weeks_reporting/52)), family = poisson(link = "log"), data = ma_tx_mumps_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -26.813   -8.256   -1.230    4.162   46.917
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.686e+02  1.598e+00  293.22   <2e-16 ***
## year        -2.417e-01  8.106e-04 -298.16   <2e-16 ***
## stateTexas   2.395e-01  7.811e-03   30.66   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 209154  on 68  degrees of freedom
## Residual deviance:  14597  on 66  degrees of freedom
## AIC: 15063
##
## Number of Fisher Scoring iterations: 5
```

```r
coefficients(test)[2]
```

13. What is the estimated yearly growth rate for Mumps in Massachusetts?

```
##       year
## -0.2416904
```

Based on this model, the estimated yearly growth rate is -24.2 percent. This means that the case rate is decreasing by 24.2 percent annually on average.

```r
coefficients(test)[2]
```

14. What is the estimated yearly growth rate for Mumps in Texas?

```
##       year
## -0.2416904
```

Based on this model, the estimated yearly growth rate is -24.2 percent. This means that the case rate is decreasing by 24.2 percent annually on average. This model is not sufficiently complex to allow for two different growth rates for Massachusetts and Texas. If you include an interaction between `state` and `year`, then it is possible to estimate a different rate for each state.

As a class, we will discuss how to interpret the results and connect this to relative risk, which is defined as

$$RR = \frac{P(\text{diseased}|\text{exposed})}{P(\text{diseased}|\text{unexposed})}.$$