Lab 3: Poisson and Logistic Regression, and ties to RR and OR

STAT 244NF: Infectious Diseases Modeling

SOLUTIONS

23 September, 2021

Infectious Disease Outbreak

An infectious disease outbreak has been reported among adults over 21 years old. Epidemiological investigators identified cases of disease over the past month and tracked possible exposures, including going to a bar within the last 10 days (1 if yes, 0 if no), and going to an outdoor park with in the last 10 days (1 if yes, 0 if no). The data for this outbreak are generated below and saved as `id_outbreak`.

1. There are two potential exposures in this fictional outbreak. What are they? The two potential exposures are going to a bar or going to an outdoor park in the last 10 days.

Calculating RR and OR based on observed data (no model)

For the following questions, you may disregard the time component.

2. Calculate the probability that a person who went to the park in the last 10 days is infected.

```r
prob.park <- id_outbreak %>%
  filter(park_last_10==1) %>%
  summarise(
    prob.park = mean(case)
  )
```

The probability that a person who went to the park in the last 10 days is infected is $P(\text{infected}|\text{park}) = 0.127$.

3. Calculate the probability that a person who did not go to the park in the last 10 days is infected.

```r
prob.nopark <- id_outbreak %>%
  filter(park_last_10==0) %>%
  summarise(
    prob.nopark = mean(case)
  )
```

The probability that a person who did not go to the park in the last 10 days is infected is $P(\text{infected}|\text{no park}) = 0.11$.

4. Calculate the probability that a person who went to a bar in the last 10 days is infected.

```r
prob.bar <- id_outbreak %>%
  filter(bar_last_10==1) %>%
  summarise(
    prob.bar = mean(case)
  )
```

The probability that a person who went to a bar in the last 10 days is infected is $P(\text{infected}|\text{bar}) = 0.341$.

```r


prob.nobar <- id_outbreak %>%
  filter(bar_last_10==0) %>%
  summarise(
    prob.nobar = mean(case)
  )
```

5. Calculate the probability that a person who did not go to a bar in the last 10 days is infected. The probability that a person who did not go to a bar in the last 10 days is infected is $P(\text{infected}|\text{no bar}) = 0.02$.

6. Calculate the relative risk, the chance that a person who went to a park will develop disease relative to the chance that a person who did not go to a park will develop disease.

```r
RR_park <- prob.park/prob.nopark
```

The relative risk of infection based on park exposure is 1.156.

Although not explicitly asked, we can do the same for bar exposure:

```r
RR_bar <- prob.bar/prob.nobar
```

The relative risk of infection based on bar exposure is 16.648.

7. Calculate the odds ratio for park exposure.

```r
OR_fcn <- function(p_e, p_u){
  odds_e <- p_e/(1-p_e)
  odds_u <- p_u/(1-p_u)
  odds_e/odds_u
}

OR_park <- OR_fcn(prob.park, prob.nopark)
```

The odds ratio of infection based on park exposure is 1.179.

8. Calculate the odds ratio for bar exposure.

```r
OR_bar <- OR_fcn(prob.bar, prob.nobar)
```

The odds ratio of infection based on park exposure is 24.741.

9. Compare the relative risk and odds ratio for bar exposure. Are they similar?

No, they are not similar (RR = 16.648 versus OR = 24.741). We only expect these to be similar under the rare disease assumption (prevalence $\leq 10\%$). We can get an estimate of the prevalence for this fictional disease by calculating the proportion of individuals in the population that have disease (regardless of exposure).

```
id_outbreak %>%
  summarise(
    prev = mean(case)
  )
```

```
##   prev
## 1 0.12
```

The prevalence is 12%, so this does not satisfy our rare disease assumption.

10. Compare the relative risk and odds ratio for park exposure. Are they similar? Yes, they are similar (RR = 1.156 versus OR = 1.179). This might seem surprising, since the rare disease assumption is violated. However, most of the source of infection (and thus most of the prevalence) is attributable to the bar exposure (where we see the large discrepancy between OR and RR), so this is OK. This gets into some of the additional nuances of using OR to estimate RR, which are beyond the scope of the class.

11. Among these two potential exposures, which is more likely to be the source of the outbreak? Why? Since the RR (or the OR) are large for bar exposure (much larger than 1), and the RR (and the OR) are close to 1 for park exposure, the bar is much more likely to be a relevant (and risky) exposure for this infection than the outdoor park. *While no claims are made about these data being biologically plausible*, this phenomenon mirrors things we know about respiratory diseases, like COVID, which was the inspiration.

Recall, if RR is close *to 1*, that means $P(\text{disease}|\text{exposure}) \approx P(\text{disease}|\text{no exposure})$. Similarly, if the OR is close *to 1*, that means the odds of disease given exposure are close to the odds of disease given no exposure. In other words the exposure under consideration is not important when it comes to spreading this disease.

Poisson regression

12. Fit a Poisson regression model with bar exposure and park exposure as explanatory (independent) variables and case as the dependent variable. Remember to use the `glm` function and to specify `family` argument in the `glm` function as `poisson`. Assign the model fit to `outbreak_pois` and print the summary of the model fit.

```
outbreak_pois = glm(data=id_outbreak, case ~ bar_last_10 + park_last_10,
                    family=poisson(link = "log"))
summary(outbreak_pois)
```

```
##
## Call:
## glm(formula = case ~ bar_last_10 + park_last_10, family = poisson(link = "log"),
##     data = id_outbreak)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8461  -0.2075  -0.2075  -0.1948   2.4428
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.9648     0.4461  -8.888  < 2e-16 ***
## bar_last_10    2.8113     0.4346   6.468  9.9e-11 ***
## park_last_10   0.1261     0.2896   0.435    0.663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 216.27  on 424  degrees of freedom
## Residual deviance: 143.32  on 422  degrees of freedom
## AIC: 251.32
##
## Number of Fisher Scoring iterations: 6
```

13. What is the estimated rate of cases, $\hat{\lambda}$, for people that did not do to a bar or the park? The equation below is a useful starting point. You will still need to solve for $\hat{\lambda}$ and calculate the estimate.

$$\log(\hat{\lambda}) = -3.9648 + 2.8113 \times 0 + 0.1261 \times 0$$

```
b <- coefficients(outbreak_pois)
rate_00 <- exp(b[1]+b[2]*0+b[3]*0)
```

The estimated rate of cases among people who did not go to a bar or park in a month (time frame for which we have data) is 0.019.

13. What is the estimated rate of cases, $\hat{\lambda}$, for people that went to a park but did not go to a bar?

```
rate_01 <- exp(b[1]+b[2]*0+b[3]*1)
```

The estimated rate of cases among people who did not go to a bar but went to a park in a month (time frame for which we have data) is 0.022.

14. In order to calculate relative risk of an exposure, we exponentiate (`exp`) the estimate of the effect of that exposure. What is the estimated relative risk of bar exposure for this disease? What is the associated 95% confidence interval?

```
RR_bar_pois <- exp(b[2])

CI_RR_bar_pois <- exp(confint(outbreak_pois, parm="bar_last_10"))
```

```
## Waiting for profiling to be done...
```

The estimated relative risk for this disease due to bar exposure is 16.631. This means that those that went to a bar in the last 10 days are 16.631 times more likely to get this disease than those that did not go to a bar in the last 10 days.

We are 95% confident that the relative risk due to bar exposure is between 7.676 and 43.493. In other words, we are 95% confident that those went to a bar in the last 10 days are between 7.676 and 43.493 times mores likely to get this disease than those that did not go to a bar in the last 10 days.

15. What is the estimated relative risk of park exposure for this disease. What is the associated 95% confidence interval?

```
RR_park_pois <- exp(b[3])

CI_RR_park_pois <- exp(confint(outbreak_pois, parm="park_last_10"))
```

```
## Waiting for profiling to be done...
```

We are 95% confident that the relative risk due to park exposure is between 0.649 and 2.037. In other words, we are 95% confident that those went to a park in the last 10 days are between 0.649 and 2.037 times mores likely to get this disease than those that did not go to a park in the last 10 days.

Binomial logistic regression

16. Fit a logistic regression model with bar exposure and park exposure as explanatory (independent) variables and count as the dependent variable. Remember to use the `glm` function and to specify `family` argument in the `glm` function as `binomial`. Assign the model fit to `outbreak_binom` and print the summary of the model fit.

```
outbreak_binom = glm(data=id_outbreak, case ~ bar_last_10 + park_last_10,
                     family=binomial(link = "logit"))
summary(outbreak_binom)
```

```
##
## Call:
## glm(formula = case ~ bar_last_10 + park_last_10, family = binomial(link = "logit"),
##     data = id_outbreak)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9399  -0.2106  -0.2106  -0.1926   2.8271
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.9777     0.4665  -8.526  < 2e-16 ***
## bar_last_10    3.2095     0.4516   7.106 1.19e-12 ***
## park_last_10   0.1800     0.3451   0.521    0.602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 311.89  on 424  degrees of freedom
## Residual deviance: 227.66  on 422  degrees of freedom
## AIC: 233.66
##
## Number of Fisher Scoring iterations: 6
```

17. What is the estimated odds of $\hat{\pi}/(1-\hat{\pi})$, for people that did not do to a bar or the park? Use the same logic as for Poisson regression. You will still need to solve for $\hat{\pi}/(1-\hat{\pi})$ and calculate the estimate.

For logistic regression, the model equation that we now have is:

$$\text{logit}(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.9777 + 3.2095 \times I(bar) + 0.1800 \times I(park)$$

where $I(bar)$ is 0 if someone did not go to a bar and 1 if they did, and $I(park)$ is 0 if someone did not go to a park and 1 if they did.

```
d <- coefficients(outbreak_binom)

odds_00 <- exp(d[1]+d[2]*0+d[3]*0)
```

The odds of disease for people who did not go to the bar or park in the last 10 days is 0.019.

18. In order to calculate odds ratio, we exponentiate the estimate of the effect of that exposure. What is the estimated odds ratio associated with bar exposure for this disease? What is the associated 95% confidence interval?

```
OR_bar <- exp(d[2])

CI_OR_bar_logit <- exp(confint(outbreak_binom, parm="bar_last_10"))
```

```
## Waiting for profiling to be done...
```

The estimated odds ratio for this disease due to bar exposure is 24.766. This means that the *odds* of getting the disease for those that went to the bar are 24.766 times the *odds* of getting the disease for those that did not go to the bar in the last 10 days.

We are 95% confident that the odds ratio due to bar exposure is between 10.992 and 66.523. In other words, we are 95% confident that the odds of getting disease for those that went to a bar in the last 10 days are between 10.992 and 66.523 times the odds of getting the disease for those that did not go to a bar in the last 10 days.

19. What is the estimated odds ratio associated with park exposure for this disease? What is the associated 95% confidence interval?

```
OR_park <- exp(d[3])

CI_OR_park_logit <- exp(confint(outbreak_binom, parm="park_last_10"))
```

```
## Waiting for profiling to be done...
```

The estimated odds ratio for this disease due to park exposure is 1.197. This means that the *odds* of getting the disease for those that went to the park are 1.197 times the *odds* of getting the disease for those that did not go to the park in the last 10 days.

We are 95% confident that the odds ratio due to park exposure is between 0.613 and 2.384. In other words, we are 95% confident that the odds of getting disease for those that went to a park in the last 10 days are between 0.613 and 2.384 times the odds of getting the disease for those that did not go to a park in the last 10 days.