Modeling Counts with the Poisson Distribution

Poisson regression – Stat 230

In this class we'll start to explore Poisson regression. To begin, we'll explore *why* we might want to use the Poisson distribution instead of the Binomial distribution. Then, we'll move on to incorporating predictor variables into our Poisson model.

Reviewing the binomial model

In these tasks, you'll explore what a simple binomial model would look like for the cancer data. As you work through these tasks, be sure to think about what the data look like, and if you think that's reasonable.

Task 1. Complete chapter 8 activity 9. For this activity, you can use the app:

https://shiny.mathcs.carleton.edu/users/aloy/simulate-binomial/

Be patient! You're running a reasonably large simulation on a slower server, so it will take a little while!

Task 2. Complete chapter 8 activity 10.

Task 3. Complete chapter 8 activities 11-12 (These questions ask you to examine Figure 8.2 in the book).

🚦 **Stop here.** We'll review the Poisson distribution as a class before moving on.

The Poisson model for count data

Another distribution that can be used to model count data is the Poisson distribution. Poisson distributions are often used to model the number of times an event occurs over an interval of time or space. Here, we're considering the number of cancer cases over a fixed level of exposure. If we denote the exposure as t and the cancer rate as θ, then we can express the Poisson probability model as

$$P(Y = y) = \frac{e^{t\theta}(t\theta)^y}{y!}, \quad \text{for} \quad y = 0, 2, 1, ...$$

The next task has you explore this new probability model and compare it to the binomial model.

Task 4. Complete chapter 8 activities 13-14.

🚦 **Stop here.** We'll discuss how to construct a Poisson regression model together as a class.

Fitting a Poisson regression model

So far we have used a rather naive model, assuming that every neighborhood will have the same probability (incidence rate) of developing cancer. However, neighborhood characteristics can impact this probability. To incorporate neighborhood characteristics into our model, we'll need a new regression model, a Poisson regression model.

Task 5. Complete chapter 8 activity 15. In R, we use the `glm()` with `family = "poisson"` to fit a Poisson regression model. Today, we are modeling rates, so we also need to include an `offset` argument. Template code for fitting this model is:

```r
glm(y ~ x, data = df, family = "poisson", offset = log(exposure))
```

where you need to replace `y`, `x`, `df`, and `exposures` with the appropriate columns/data frame names.

Task 6. Complete chapter 8 activity 16.