Overview

Submit your writeup including any code and plots as a PDF via Gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to reuse code will help in the long run!

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

1. GLM for Dilution Assay

Being able to reformulate problems as generalized linear models (GLMs) enables you solve a wide variety of problems with existing packages. We recommend reviewing the examples of GLMs from Lectures 10 and 11. In particular, make sure you understand that formulating a GLM involves choosing an 1) output distribution and 2) link function that are appropriate for the application at hand.

In this problem, you'll retrace the footsteps of the statistician R. A. Fisher and develop one of the very first applications of GLMs. In a 1922 paper, Fisher formulated a GLM he used to estimate the unknown concentration ρ_0 of an infectious microbe in a solution. Without specialized technology to directly measure ρ_0 from the solution, Fisher devised the following procedure: we will progressively dilute the original solution, and after each dilution, we'll pour out some small (fixed) volume v onto a sterile plate. If zero microbes land on the plate, it will remain sterile, but if any microbes land on a plate, they will grow visibly on it (we call this an "infected plate"). Eventually, the solution should become diluted enough that no microbes will land on the plate. By observing whether or not the plate is infected at each dilution, and by formulating the relationship between this data and ρ_0 as a GLM, we can estimate ρ_0 from this data.

Specifically, let ρ_t denote the concentration at dilution t. Each time, we dilute the solution to be half its concentration, such that

$$\rho_t = \frac{\rho_0}{2^t} \tag{1}$$

for $t = 0, 1, \ldots$. When we pour out volume v of the solution onto the plate, and wait awhile to allow for microbe growth, we can observe whether a plate was infected (*i.e.*, has a non-zero number of microbes) or is sterile (*i.e.*, has zero microbes). Therefore, our data $Y_t \in \{0, 1\}$ is whether or not the plate is infected at each dilution.

We'll formulate a GLM that relates ρ_0 and t to the data Y_t. Estimating the parameters of this GLM will then allow us to estimate ρ_0, as will become clear in the last part.

[1] In Jupyter, you can download as PDF or print to save as PDF

We'll start by choosing an output distribution (part (a)). Then, we'll use our knowledge of how the microbes grow to choose a link function (parts (b) and (c)). Finally, you'll write out a formula you can use to predict the initial concentration ρ_0 based on the input data and the model coefficients (part (d)).

(a) (2 points) At dilution t, the data $Y_t \in \{0, 1\}$ indicates whether or not the plate is infected. The chance that a plate gets infected is denoted by $\mu(t) := \mathbb{E}[Y_t]$. Write down an output distribution for Y_t that is appropriate for the values it takes on, using $\mu(t)$ as a parameter (we'll derive what $\mu(t)$ should be in the next part).

(b) (5 points) At dilution t, we pour out volume v onto a plate, so the expected number of microbes on the plate is $\rho_t v$. The actual number of microbes is distributed as a Poisson random variable with this mean $\rho_t v$:

$$\#\text{ microbes on plate at dilution } t \sim \text{Poisson}(\rho_t v). \tag{2}$$

Using this fact, write out an expression for $\mu(t) := \mathbb{E}[Y_t]$. Start with

$$\begin{aligned} \mu(t) &= \mathbb{P}(\text{plate is infected at dilution } t) && (3) \\ &= 1 - \mathbb{P}(\text{there are 0 microbes on plate at dilution } t). && (4) \end{aligned}$$

(c) (5 points) Use your findings from part (b), along with Equation (1), to find a link function g such that

$$g(\mu(t)) = \beta_0 + \beta_1 t \tag{5}$$

for some constants β_0 and β_1. (Remember that in class, we talked about the inverse link function g^{-1}, such that $\mu(t) = g^{-1}(\beta_0 + \beta_1 t)$).

(d) (3 points) Choosing an appropriate output distribution and link function as we've done in Parts (a) and (c) completes the GLM specification. Now, suppose you've estimated β_0 and β_1 (*e.g.*, using maximum-likelihood estimation). Write down an estimate of ρ_0.