Generalized Linear Models

Count Data

Demetris Athienitis

Motivation

Sometimes our data are in the form of counts

- Number of crimes in a particular region
- Number of games a team will win

To model these data, we first need a different distribution...the Poisson.

Poisson distribution

The PMF is

$$p(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, \ldots \text{ and } \mu > 0$$

with $E(Y) = V(Y) = \mu$

The mean and variance being the same can sometimes be too restrictive. Will deal with this later.

Section 2

GLM on Counts

The default link function for count data is the log link (called log-linear modeling), which ensures μ is positive

$$\begin{aligned}\log(\mu) &= \alpha + \beta x \\ \Rightarrow \mu &= e^{(\alpha+\beta x)} \\ &= e^{\alpha}(e^{\beta})^{x}\end{aligned}$$

Example

- Y = number of defects of silicon wafer
- $x = 0$ if type A, 1 if type B

A	8	7	6	6	3	4	7	2	3	4
B	9	9	8	14	8	13	11	5	7	6

Fit log-linear model to see if mean defect number depends on group

```
> wafers.log=glm(defects~trt,family=poisson(link="log"),
+  data=wafers)
> summary(wafers.log)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
trtB          0.5878     0.1764   3.332 0.000861 ***
```

Example (continued)

$$\log\left[\mu(x)\right] = 1.6094 + 0.5878x$$

A: $\mu(0) = \exp(1.6094) = 5$

B: $\mu(1) = \exp(1.6094)\exp(0.5878) = 5 + 4 = 9$

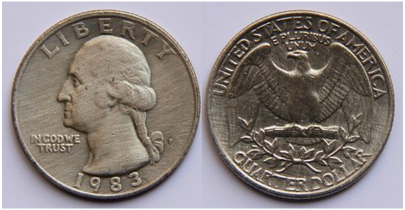
And a 95% CI on (α and) β does not include 0

```
> confint(wafers.log)
                2.5 %    97.5 %
(Intercept) 1.3188383 1.8743819
trtB        0.2469096 0.9400962
```

Group status does have a significant effect on defects, with B being larger.

Labeling

- What if $x = 1$ if type A, 0 if type B, i.e. switch labels. Would the conclusions differ?
- Which coin is a US 25 cent coin (quarter)?



BOTH! Just a matter of perspective.

Section 3

Count data with different bases

Sometimes count data have different bases.

Example

Imagine modeling the number of COVID-19 cases in Gainesville and Atlantay

- Will appear that ATL has FAR more cases than GNV
- ATL simply has more people and therefore more crimes
- We are more interested in the positivity rates between the two cities
- How many cases are there per capita

Count data with different bases

Example

Person A catches 11 fish, and person B catches 20. Who is the better fisherman?

Person A

Person B

- Have to account for the net sizes
- Person A is actually pretty impressive

GLMs for count data

- Let Y be the count and t be the base

$$E\left(\frac{Y}{t}\right) = \frac{\mu}{t}$$

- Hence, we can do the following

$$\begin{aligned} & \log\left(\frac{\mu}{t}\right) = \log(\mu) - \log(t) = \alpha + \beta x \\ \Rightarrow & \log(\mu) = \alpha + \beta x + \log(t) \\ \Rightarrow & \log(\mu) = \alpha + \beta x + \underbrace{\beta_2}_{=1} \underbrace{x_2}_{\log(t)} \end{aligned}$$

- Add another "predictor" whose coefficient is set to 1, the term $\log(t)$ is called the *offset*

Example

Data on the number of airline deaths between 1995 and 2017

Fatalities	Available seat miles	Year
1828	829581	1995
2796	862621	1996
1768	884192	1997
1721	898359	1998
1150	945245	1999
⋮	⋮	⋮

- Interested in the rate of fatalities per seat mile
- Will use an offset term for the number of seat miles

Example (continued)

```
> air.poisson=glm(Fatalities~I(Year-1995),family=poisson,
+  data=air_deaths, offset=log(ASM))
> summary(air.poisson)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -6.0541485  0.0101474 -596.62   <2e-16 ***
I(Year - 1995) -0.0638961  0.0009377  -68.14   <2e-16 ***
```

$$\log\left(\frac{\hat{\mu}}{t}\right) = -6.05 - 0.06 \times (\text{Year} - 1995)$$

$$\Rightarrow \frac{\hat{\mu}}{t} = e^{-6.05} e^{-0.06 \times (\text{Year} - 1995)}$$

Example (continued)

- β is significantly different from 0 (with small p-value)
- The rate appears to be going down over time
- Each year the rate is $e^{-0.06} = 0.94$ times what it was the previous year

Example

British train collision example provided in class notes

We learned

- For Count Data, we use Possion distribution with log link
- If necessary take into account different bases