Poisson Regression

Motivation and Set-up:

In order to model the expected *count* or rate of some process, we use a Poisson model.

- Disease occurrence
- Claims made (Insurance)

If *exposure* is constant across all subjects, the mean count can be modeled directly. However usually the **rate** is modeled. E.g. cases / 100 patient-years or events / month, etc.

Set-up: for $i = 1, ..., n$ i.i.d samples

- Y_i = event count
- $\boldsymbol{x}_i$ = covariate vector
- T_i = exposure; $\log T_i$ is the offset
- rate $= \frac{E[\text{count}]}{\text{exposure}} = \frac{\mu_i}{T_i} = \lambda_i$
- $Y_i \sim$ Poisson $(\lambda_i = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})$
- $P(Y_i | \boldsymbol{x}_i, T_i) = \frac{e^{\lambda_i} \lambda_i^{Y_i}}{y_i!}$

Rate Model

- $\log(\lambda_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- $\log E[\frac{y_i}{T_i}] = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- $\log \underbrace{E[y_i]}_{\mu_i} = \log T_i + \boldsymbol{x}_i^T \boldsymbol{\beta}$

GLM details:

- log link; $\log \lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- Mean function:

  $\mu_i = T_i \lambda_i \iff \frac{\mu_i}{T_i} = \lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- Variance function $v(\mu_i) = \mu_i = T_i \lambda_i$
- $a(\phi) = 1$ – usually, check dispersion section

Interpretation

Consider the following toy example

$$\log(\lambda_i) = \beta_0 + \beta + 1(X_i - \bar{X})$$

Coefficient interpretations

- β_0

  - if $x_i = \bar{x}$ then $\beta_0 = \log(\frac{\mu_i}{T_i}) = \log(\lambda_i)$

  - so β_0 is the log rate when $x_i = \bar{x}$
- β_1

  $$\begin{aligned} \log \lambda_a - \log \lambda_b &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1(x) \\ &= \log \frac{\lambda_a}{\lambda_b} = \beta_1 \end{aligned}$$

  - So β_1 is the log rate ratio for a one unit increase in x

Likelihood Function and Derivatives

$$
\begin{aligned}
L(\boldsymbol{\beta}) &\propto \prod_{i=1}^{n} \frac{e^{\lambda_i} \lambda_i^{Y_i}}{Y_i!} \\
&\propto \prod_{i=1}^{n} \frac{\exp\{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\} \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}^{Y_i}}{Y_i!} \\
\Rightarrow l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} -T_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + Y_i \log T_i + Y_i \boldsymbol{x}_i^T \boldsymbol{\beta} \\
U(\boldsymbol{\beta}) &= \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - T_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) \\
J(\boldsymbol{\beta}) &= \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T T_i \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\} \\
\boldsymbol{V} &= \operatorname{diag}(v(\mu_i)) = \operatorname{diag}(T_i e^{\boldsymbol{x}_i^T \boldsymbol{\beta}})
\end{aligned}
$$

Poisson Regression Example

Context: Analyzing the coronary heart disease of 3,154 males aged 40-50 in a prospective cohort design. We're interested in modeling the number of CHD cases. Risk factors recorded include smoking, blood pressure and behavior type (A and B).

We first fit a main effects model without an offset.

- lack of offset ruins interpretation

  not all men had same exposure

  if all men *had* had the same exposure, this model would be fine
- including offset restores interpretative value

- Changing the exposure from Person Years to Person Years / 1000 only changes the **intercept**

  $$\beta_0 = \beta_0^* - \log(1000)$$

  $$\beta_0^* = \beta_0 + \log(1000)$$