Count Data

The number of times an event occurs is a common form of data.

Examples include

- # of damages caused by waves to vessels
  - covariates: ship type, year of construction, period of operation, aggregated month in service;
  - interested in modelling the relationship between the average number of damages and the covariates.
- # of breast cancer cases in Iceland during 1910-1920 in certain age group
  - demographics are covariates

Poisson distribution

Suppose $Y \sim Poisson(\lambda)$, then

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, y = 0, 1, 2, \ldots.$$

- Mean: $\mu = \lambda$
- Variance: $V(\mu) = \mu = \lambda$

Poisson Regression

- The log likelihood function of $y_i \sim Poisson(\lambda_i)$ is

$$l(\theta_i, y_i) = (y_i\theta_i - e^{\theta_i}) + c,$$

  where the canonical parameter $\theta_i = \log \lambda_i$.

- Set $\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. We obtain the Poisson log linear model (with the canonical log link function).

- Deviance:

$$D(\mathbf{Y}; \hat{\mu}) = 2 \sum_{i=1}^{n} \{ Y_i \log \frac{Y_i}{\hat{\mu}_i} - (Y_i - \hat{\mu}_i) \}.$$

- Pearson χ^2:

$$G = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Both are approximately $\chi^2(n-p)$ (if μ_is are large).

Poisson Rate and Offset

When events occur independently, successively and at the *same rate*, the Poisson distribution is appropriate for the # of events observed.

- Usually the expectation is a product of the *Poisson rate* and length of the observing period, or *"exposure"*.
- Example: The number car accidents on GWB follows a Poisson distribution, with a constant Poisson rate λ. The exposure is the number of days n of each observation period.
- Example: suppose Y_i is the number of insurance claims for a particular make/model of car. This depends on the number of insured cars of this type, n_i, and other variables affecting λ_i such as length of warranty and manufacture location of the car.

Let Y_i denote the number of events observed from exposure n_i for the ith covariate pattern, and they are independent for $i = 1, \ldots, N$.

- Expectation of Y_i is
$$\mathbb{E}(Y_i) = \mu_i = n_i \lambda_i.$$
- In GLM, we care more about λ_i rather than μ_i

Model: The Poisson log linear model with offset is

$$\begin{cases} Y_i \sim Poisson(\mu_i), \\ \mathbb{E}(Y_i) = \mu_i = n_i \lambda_i = n_i \exp(\boldsymbol{x}_i \boldsymbol{\beta}). \end{cases}$$

with the canonical link function,

$$\log \mu_i = \log n_i + \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

- The term $\log n_i$ is called the offset.
- It is a known constant, which is readily incorporated into the estimation procedure.
- Parameter estimates are usually interpreted in terms of log rate ratio.

Goodness of Fit

The fitted values are given by

$$\hat{y}_i = \hat{\mu}_i = n_i \exp(\boldsymbol{x}_i \hat{\boldsymbol{\beta}}).$$

The Pearson residual is

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

The deviance residual is

$$d_i = sign(y_i - \hat{\mu}_i)\sqrt{2[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]}.$$

The two goodness-of-fit statistics are

$$X^2 = \sum r_i^2 \quad \text{and} \quad D = \sum d_i^2.$$

Example: Wave Damage

It is of interest to investigate the risk of damage associated with factors of ship type, year of construction, and period of operation.

```
#   Type Year   Period Month    Damage
#1     A 60-64  60-74   127       0
#2     A 60-64  75-79    63       0
#3     A 65-69  60-74  1095       3
#4     A 65-69  75-79  1095       4
#5     A 70-74  60-74  1512       6
```

We consider two models:

- Model 1: Type+Year+Period
- Model 2: Type+Year+Period+Year*Type