Count data

- Refers to number of times and events. Frequency data.
- "Number of tornados per month", "hurricanes per year".
- Often modeled with a Poisson distribution of parameter λ.

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}; y = 0, 1, 2, \ldots,$$

- λ is the average number of ocurrances.
- **Poisson regression:** $Y_1, Y_2, \ldots, Y_n$ are n counts.
- where Y_i denotes the number of events for "exposures" η_i.

- So $E(Y_i) = \eta_i \theta_i$ and observation i has a specific covariance pattern.
- θ_i is explained through covariates,

$$\theta_i = exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}) = exp(\mathbf{x}_i^t \beta).$$

- The model is

$$Y_i \sim Poisson(\lambda_i); \quad \lambda_i = \eta_i \theta_i = \eta_i exp(\mathbf{x}_i^t \beta), i = 1, 2, \ldots, n$$

- In log-scale,

$$\begin{aligned} log(\lambda_i) &= log(\eta_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \\ &= \text{offset} + \text{linear predictor} \end{aligned}$$

- For a covariate X_j, factor is absent $X_j = 0$ and factor present if $X_j = 1$.
- The *Rate ratio* (RR)

$$RR = \frac{E(Y_i|\text{present})}{E(Y_i|\text{absent})} = \frac{\eta_i exp(\beta_0 + \beta_1)}{\eta_i exp(\beta_0)} = exp(\beta_1)$$

- If a covariate is increased by one unit, $exp(\beta_1)$ is the effect due to the increase.
- In general, the RR for covariate i is $exp(\hat{\beta}_i)$.
- An approximate 95 % confidence interval for the RR is

$$(\exp(\hat{\beta}_i - 1.96SE(\hat{\beta}_i)), \exp(\hat{\beta}_i + 1.96SE(\hat{\beta}_i)))$$

UNM

- *Fitted values* are computed as

$$\hat{Y}_i = \eta_i exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots + \hat{\beta}_p X_{ip})$$

  also denoted as e_i (expected values).
- These provide estimates of λ_i for each observation i, $(\hat{\lambda}_i)$.
- *Pearson residuals*,

$$r_i = \frac{Y_i - e_i}{\sqrt{e_i}}; i = 1, 2, \ldots, n$$

- Goodness of fit statistic,

$$X^2 = \sum_i r_i^2 = \sum_i \frac{(Y_i - e_i)^2}{e_i}$$

Deviance statistic for Poisson model

- *Saturated model*. Model where all λ_is are different.
- The MLE is $\hat{\lambda}_i = Y_i$.
- The maximum value of the log-likelihood is

$$l(b_{max}; Y) = \sum_i y_i log(y_i) - \sum_i y_i - \sum_i log(y_i!)$$

- For a model with $p < n$ parameters, $\hat{\beta}$ induces $\hat{\lambda}_i = \hat{Y}_i$.
- The maximum log-likelihood value is

$$l(b; Y) = \sum_i y_i log(\hat{y}_i) - \sum_i \hat{y}_i - \sum_i log(y_i!)$$

- The *deviance* is, $D = -2[l(b; Y) - l(b_{max}; Y)]$ or

$$\begin{aligned} D &= 2\left[\sum_i y_i log(y_i/\hat{y}_i) - \sum_i (y_i - \hat{y}_i)\right] \\ &= 2\sum_i [o_i log(o_i/e_i) - (o_i - e_i)] \\ &= 2\sum_i o_i log(o_i/e_i). \end{aligned}$$

- Since for most cases $\sum_i o_i = \sum_i e_i$ (model with β_0).
- The *deviance residuals* are defined as

$$d_i = \text{sign}(o_i - e_i)\sqrt{[o_i log(o_i/e_i) - (o_i - e_i)]}$$

and so

$$D = \sum_i d_i^2.$$

- With a first order Taylor series approximation,

$$o \, log\left(\frac{o}{e}\right) \approx (o - e) + \frac{1}{2}\frac{(o - e)^2}{e}$$

- Therefore,

$$\begin{aligned} D &= 2\sum_i o_i \left(\frac{o_i}{e_i}\right) \approx 2\left[\sum_i (o_i - e_i) + \frac{1}{2}\sum_i \frac{(o_i - e_i)^2}{e_i}\right] \\ &= \sum_i \frac{(o_i - e_i)^2}{e_i} = X^2 \end{aligned}$$

- Shows that D and X^2 are closely related.
- Gets compare with $\chi^2_{(n-p)}$ where $p =$ number of fitted parameters.

- **Minimal model** with no covariates,

$$log(\lambda_i) = log(\eta_i) + \beta_0$$

- If $l(b_{min})$ is the maximum likelihood under this model.
- $l(b)$ the max. likelihood of a model with p parameters.
- *Likelihood chi-square statistic*,

$$C = 2[l(b) - l(b_{min})]$$

- *pseudo* R^2 measure is

$$R^2 = \frac{l(b_{min}) - l(b)}{l(b_{min})}.$$

- *Example (Table 9.1):* Number of deaths from coronary heart disease.
- Total number of person-years of observations (offset) .
- Covariates: Age group, smokers/non-smokers.
- Death rate higher for smoker and non-smokers?
- Is death rate related to Age?
- A model is,

$$\begin{aligned} log(deaths_i) &= log(pop_i) + \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i \\ &+ \beta_4 agesq_i + \beta_5 smkage_i; i = 1, \ldots, 10 \end{aligned}$$

- $agesq_i$ is the square of *agecat*
- *smkage* is the *smoke* and *age* interaction.