Stat 574: GLMs and Survival Analysis

Gabriel Huerta

Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM, E.U.A.

Chapter 1, summary notes.

Poisson regression example

- N_i number of virus mutations for patient $i = 1, 2, \ldots, n$.
- T_i is the time between treatment visits.
- Model: $N_i \sim Poisson(e^{\beta_0 + \beta_1 T_i}); i = 1, 2, \ldots, n$
- Rate parameter is $\lambda_i = e^{\beta_0 + \beta_1 T_i}$.
- The mean response $E(N_i) = \lambda_i$ and $\lambda_i > 0$

$$log(E(N_i)) = log(\lambda_i) = \beta_0 + \beta_1 T_i$$

- a function of the mean gives us a *Linear model* (in this case log). **Link function**
- What is the link function in *linear regression*?

Ideas on Survival Analysis

- Observed time: $T_i, i = 1, 2, \ldots, n$ (time to event).
- Covariates: $X_{ij}; i = 1, 2, \ldots, n, j = 1, \ldots, p$ (factors that relate to ”failure).
- Censoring variable: $\delta_i; i = 1, 2, \ldots, n$ (people move or leave. no follow-up).
- Difficult part: incorporate δ into the *likelihood* function.
- *Kaplan-Meier* estimator of CDF of T to account for biases.
- Parametric Model: Exponential, Weibull or Gumbel.
- *Cox proportional hazard* model: semi-parametric. incorporates X_{ij}.

Maximum Likelihood Estimation (MLE)

- $Y = [Y_1, Y_2, \ldots, Y_n]^T$. Joint probability distribution $f(Y|\theta)$.
- *Likelihood function:* $Y = y$ is observed (fixed). θ unknown.
- Ω parameter space; set of all possibilities for θ.

$$L(\theta|Y = y) = f(y|\theta)$$

- MLE is the value θ, $\hat{\theta}$, that maximizes $L(\theta|Y = y)$.
- So $log(L(\hat{\theta}|Y = y)) \geq log(L(\theta|Y = y))$ for all $\theta \in \Omega$.
- If θ is a scalar, try to solve

$$\frac{dlog(L(\theta|Y = y))}{d\theta} = 0$$

- In GLM's and Survival models $\hat{\theta}$ is obtained numerically (Newton-Raphson).

Example: Tropical cyclones

- Y_i number of tropical cyclones in successive seasons. $n = 13$.
- Data: $6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4$.
- Model Y_i as independent variables $Y_i \sim Poisson(\theta)$.

$$f(Y_i|\theta) = \frac{\theta^{Y_i} e^{-\theta}}{y_i!}; y_i = 0, 1, 2, \ldots$$

- To find MLE or $\hat{\theta}$,
  - Find joint distribution $f(y_1, y_2, \ldots, y_n|\theta)$
  - Take 'log' and find first derivative with respect to θ.
- *Likelihood equation*

$$\frac{dl}{d\theta} = \frac{\sum Y_i}{\theta} - n = 0.$$

- Solve for θ implies

$$\hat{\theta} = \frac{\sum Y_i}{n} = \bar{Y} = 5.538$$

- Check second derivative is negative at $\hat{\theta}$.
- **Teaser on Bayes**: In addition to $f(y_1, y_2, \ldots, y_n|\theta)$ we need a *prior* on θ, $p(\theta)$ or $\pi(\theta)$.
- $p(\theta)$ represents our 'state of uncertainty' of θ.
- In this example, one could use a $Gamma(a, b)$ prob. distribution.
- *Bayes theorem:* provides the 'posterior distribution',

$$p(\theta|y_1, y_2, \ldots, y_n)\ (\textit{prop to prior} \ \times \ \textit{likelihood})$$

- requires probability distributions, solving integrals or Openbugs/Winbugs.

Least Squares

- $Y_1, Y_2, \ldots, Y_n$ with means $\mu_1, \mu_2, \ldots, \mu_n$ (cyclone example $\mu_i = \theta_i$).
- Suppose each μ_i is a function of $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$) so $E(Y_i) = \mu_i(\beta)$.
- *Least squares:* Find $\hat{\beta}$ that minimizes

$$S = \sum_{i=1}^{n} (Y_i - \mu_i(\beta))^2 (\mathit{sum\ of\ errors\ squared})$$

- Simultaneously solve $\frac{dS}{d\beta_j} = 0; j = 1, 2, \ldots, p$.
- For the cyclone data example, $\theta = e^\beta$ (or $log(\theta) = \beta$).
- The sum of squares is $S = \sum_i (Y_i - e^\beta)^2$. $\hat{\beta}$?

- **Weighted Least Squares**: Suppose $Var(Y_i) = \sigma_i^2$ (depends on observations).
- Minimize

$$S = \sum_{i=1}^{n} w_i [Y_i - \mu_i(\beta)]^2$$

- where $w_i = 1/\sigma_i^2$ (reciprocal variance).
- 'observations with large variance have less influence on estimates'.
- **Distributional results:** (Section 1.4 Dobson's book)
  - Linear combinations of Normal variables.
  - chi-square (sum of) distribution.
  - t-distribution (from N(0,1) and chi-square).
  - F-distribution.

UNM