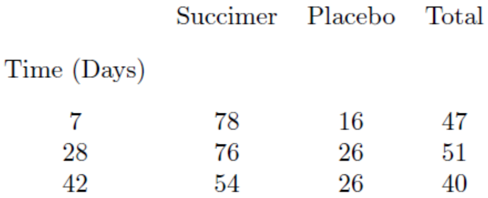
BIOS 755: Generalized Linear Models I

Alexander McLain

February 20, 2024

A Motivation Example

- Recall a previous example: Treatment of Lead-Exposed Children Trial, 100 children were randomized equally to succimer and placebo.
- The percentages of children with blood lead levels below 20 $\mu g/dL$ at the three examinations after treatment were as follows:



	Succimer	Placebo	Total
Time (Days)			
7	78	16	47
28	76	26	51
42	54	26	40

A Motivation Example

- Question: How can we quantify the effect of treatment with succimer on the probability of having a blood lead level below 20 $\mu g/dL$ at each occasion?
- Question: How can we test the hypothesis that succimer has no effect on these probabilities?
- If we had observations at only a single time point, we could model the relative odds using logistic regression.
- Here, we have to carefully consider the goals of the analysis and deal with the problem of correlation among the repeated observations.
- Before we do this, we consider just the ordinary regression situation where responses are scalar and independent (i.e. cross-sectional data).

Logistic Regression

- We first will discuss the logistic regression model for a single response variable.
- Let Y be a binary response where Y=1 represents a ‘success’ and Y=0 represents a ‘failure.’
- The mean of the binary response variable, denoted by π, is the proportion of successes or the probability that the response takes on the value 1.

$$\pi = E(Y) = Pr(Y = 1) = Pr(`success')$$

- With a binary response, we are usually interested in estimating the probability π and relating it to a set of covariates.

Logistic Regression

- A naive strategy for modeling a binary response is to consider a linear regression model
$$\pi = E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$
- However, in general, this model is not feasible since π is a probability and is restricted to values between 0 and 1.
- Also, the usual assumption of homogeneity of variance would be violated since the variance of a binary response depends on the mean

$$var(Y) = \pi(1 - \pi)$$

Logistic Regression

- Instead, we can consider a logistic regression model where

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- This model accommodates the constraint that π is restricted to values between 0 and 1.
- $\pi/(1-\pi)$ is defined as the odds of success.
- Therefore, modeling with a logistic function can be considered equivalent to a linear regression model where the mean of the continuous response has been replaced by the logarithm of the odds of success.
- Note that the relationship between π and the covariates is non-linear.

Logistic Regression

- We can use ML estimation to obtain estimates of the logistic regression parameters, assuming that the binary responses are Bernoulli random variables.
- Given the logistic regression model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- The population intercept, β_0, has interpretation as the log odds of success when all of the covariates take on the value zero.
- The population slope, say β_1, has interpretation in terms of the change in log odds of success for a single-unit change in X_1 given that all of the other covariates remain constant.

Logistic Regression

- When one of the covariates is dichotomous, say X_1, then β_1 has a special interpretation:

  $\exp(\beta_1)$ *is the odds ratio or ratio of odds of success for the two possible levels of* X_1 *(given that all other covariates remain constant).*

- Odds ratios can be directly calculated from case-control studies, where the sampling strategy focuses on the outcome rather than the exposure.
  - This is because the odds ratio does not depend on the actual prevalence of the outcome in the population, which is unknown in case-control designs.
- Relative risk, on the other hand, requires information on the actual risk in both exposed and unexposed groups, which is typically available from cohort studies but not from case-control studies.

Logistic Regression

- In studies of rare events, the odds ratio can approximate the relative risk. This is particularly useful because the odds of the event (the ratio of events to non-events) can be a more stable estimate when the event is rare, avoiding the potential for extremely large or undefined relative risk values.
- However, odds ratio has its drawbacks (see here and here).
    - The odds ratio has the disadvantage of ignoring the level, ie, the ratio 1:10 is the same as 10:100.
    - OR is good for establishing causal relations but is not that useful to the public health practitioner who is interested in knowing how much decrease in disease burden will be achieved by specific interventions. RR is a better measure than OR for such public health purposes.

Interaction

- In a multiple logistic model, the interaction “between X_1 and X_2” must also involve the probability that $Y = 1$.
- This means that the relationship between the probability that $Y = 1$ and X_1 differs (is modified) for different values of X_2.
- Interaction can be called effect modification.
- The interaction model can be written as:

$$logit\{\pi(\boldsymbol{X})\} = \log\left\{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_{12} X_1 X_2$$

where $\boldsymbol{X} = \{X_1, X_2\}$.

Interpretation with Interaction

- Suppose that X_1 is a binary variable where $X_1 = 1$ if the person is a smoker, and $X_1 = 0$ otherwise, and $X_2 = 1$ if the person is female.
- Let Y be the indicator of CHD. We fit the model

$$logit\{\pi\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_{12} X_1 X_2$$

- What is the probability of CHD for a:
  1. a smoking female,
  2. a non-smoking female,
  3. a smoking male, and
  4. a non-smoking male.
- What is the OR for CHD when comparing:
  1. a smoking female to a smoking male, and
  2. a smoking female to a non-smoking male.

Interpretation with Interaction

- With no interaction the odds ratio associated with a one unit change in X_1 is e^{β_1}.
- Let $\boldsymbol{X} = \{X_1, X_2\}$ and $\boldsymbol{X}^* = \{X_1 + 1, X_2\}$. When the interaction is included we have

$$\frac{\text{Odds for } \boldsymbol{X}^*}{\text{Odds for } \boldsymbol{X}} = \frac{\frac{\pi(\boldsymbol{X}^*)}{1-\pi(\boldsymbol{X}^*)}}{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}} \quad = \quad \frac{e^{\beta_0+\beta_1(X_1+1)+\beta_2X_2+\gamma_{12}(X_1+1)X_2}}{e^{\beta_0+\beta_1X_1+\beta_2X_2+\gamma_{12}X_1X_2}}$$

$$= \quad e^{\beta_1+\gamma_{12}X_2}$$

- In general, if $\boldsymbol{X} = \{X_1, X_2\}$ and $\boldsymbol{X}^* = \{X_1 + \Delta, X_2\}$ then the OR of $\boldsymbol{X}^*$ in reference to $\boldsymbol{X}$ is

$$e^{\beta_1\Delta+\gamma_{12}\Delta X_2}$$

Logistic Regression

- To fit logistic regression models in SAS, we can use proc logistic.
- SAS code:

```
proc logistic data=xxx;
model y=x;
run;

proc genmod data=xxx;
model y=x/ link=logit dist=bin;
run;
```

Poisson Regression

- In Poisson regression, the response variable is a count (e.g., the number of cases of a disease in a given period of time), and the Poisson distribution provides the basis of likelihood-based inference.
- Often the counts may be expressed as rates when the count or absolute number of events is often not satisfactory for comparison.
  - Why would that be?
- Like a proportion or probability, a rate provides a basis for direct comparison.
- In either case, Poisson regression relates the expected counts or rates to a set of covariates.

Example

- We want to model the rate of skin cancer as a function of location and age group.
- We have observed counts of skin cancer in two locations.

Age	Minn.-St. Paul		Dallas-Ft. Worth	
Group	Cases (Y)	Pop. (N_i)	Cases (Y)	Pop. (N_i)
15-24	1	172675	4	181343
25-34	16	123065	38	146207
35-44	30	96216	119	121374
45-54	71	92051	221	111353
55-64	102	72159	259	83004
65-74	130	54722	310	55932
75-84	133	32185	226	29007
85+	40	8328	65	7538

Poisson Regression

- The response variable is a count and is assumed to have a Poisson distribution. That is, the probability a specific number of events, Y, occurs is

  $$\Pr(Y = y) = e^{-y}\lambda^{y}/y!$$

  where λ is the expected count or number of events
- Commonly, the count is relative to a time period or population size. For example,
  - number of seizures in a two-week period, or
  - number of COVID cases per state.

  Whenever you have a count, always think about what it is relative to.

Poisson Regression

- Let t be what the count is relative to.
- The expected rate per unit t is modeled as

$$\log(\lambda/t) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- Note that since $\log(\lambda/t) = \log(\lambda) - \log(t)$, the Poisson regression model can also be considered as

$$\log(\lambda) = \log(t) + \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

where the 'coefficient' associated with $\log(t)$ is fixed to be 1. **This adjustment term is known as an 'offset' (more on this later).**

Poisson Regression

- Given the Poisson Regression model

$$\log(\lambda/t) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- The population intercept, β_0, has interpretation as the log expected rate when all the covariates take on the value zero.
- The population slope, say β_1, has interpretation in terms of the change in log expected rate for a single-unit change in X_1 given that all of the other covariates remain constant.
- When one of the covariates is dichotomous, say X_1, then β_1 has a special interpretation: $\exp(\beta_1)$ is the rate ratio for the two possible levels of X_1 (given that all of the other covariates remain constant).

Offset

- When "t" varies by person, we need to adjust for it using an offset variable, e.g., the population.
- The Poisson regression model will then be modeling the event rate per unit exposure
  - e.g., N_i: population in area i; λ_i: the rate for area i.
    We are interested in modeling the rate per person in area i:

$$\begin{aligned} log(\lambda_i/N_i) &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \\ log(\lambda_i) - log(N_i) &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \\ log(\lambda_i) &= log(N_i) + \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \end{aligned}$$

  - $\log(N_i)$ is the offset (which can be specified in `SAS` using `offset=`)

Offset

- The log of the population will be the offset
- This can be modeled in SAS with

  ```
  proc genmod data=...;
  class city agegr;
  model cases = city agegr/d=p link=log type3 offset=lpop;
  run;
  ```

  where `lpop` is defined as $\log(N_i)$ in the data statement.

Introduction to Generalized Linear Models

Generalized Linear Models

- The generalized linear model is actually a family of probability models that includes the normal, Bernoulli, Poisson, and Gamma distributions.
- Generalized linear models extend the methods of regression analysis to settings where the outcome variable can be a dichotomous (binary) variable, an ordered categorical variable, or a count.
- The generalized linear model has some of the properties of the linear model.
- Most importantly, a parameter related to the expected value is assumed to depend on a linear function of the covariates.

Generalized Linear Models

- Let Y_i, $i = 1, \ldots, n$, be independent observations from a probability distribution that belongs to the family of statistical models known as generalized linear models.
- The probability model for Y_i has a three-part specification:
  - The formula
  - The distributional assumption
  - The link function

The formula

- Given covariates $X_{i1}, \ldots, X_{ik}$, the effect of the covariates on the expected value of Y_i is expressed through the **linear predictor**

$$\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} = \boldsymbol{X}_i \boldsymbol{\beta}$$

- Although it is a 'linear' predictor (i.e. a sum), in general, this predictor can contain non-lear quantities (i.e., t^2 or t^3).

The Distributional Assumption

- Y_i is assumed to have a probability distribution that belongs to the exponential family, which includes:
  - **Continuous data:** the normal distribution, the gamma distribution (not common), inverse Gaussian (not common).
  - **Zero/one data:** Bernoulli distribution
  - **Categorical data:** multinomial
  - **Count data:** Poisson, negative binomial distributions.
- Some of these have multiple links that can be used, others tend to use the same link.

The Link Function

- The link function, $g(\cdot)$, describes the relation between the linear predictor and the expected value of Y_i (denoted by μ_i),

$$g\left\{E(Y_i)\right\} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} = \boldsymbol{X}_i \boldsymbol{\beta}$$

$$E(Y_i) = g^{-1}(\boldsymbol{X}_i \boldsymbol{\beta})$$

  where g^{-1} is the inverse of g.

- For example, for the logit link we have

$$g(\mu_i) = \log\left\{\frac{\mu_i}{1-\mu_i}\right\} \text{ then } \quad g^{-1}(X_i\boldsymbol{\beta}) = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}.$$

Common Link Functions

- **Count variables:** log-link function. $e^x > 0$ for all x.
- **Zero/One variables:** since the mean μ_i is π_i, with $0 < \pi_i < 1$, we would prefer a link function that transforms the interval $[0, 1]$ on to $[-\infty, +\infty]$.

$$\begin{aligned} \text{logit:} \quad g(\pi) &= \log\{\pi/(1-\pi)\} \\ \text{probit:} \quad g(\pi) &= \Phi^{-1}(\pi) \\ \text{complementary log-log:} \quad g(\pi) &= \log\{-\log(1-\pi)\} \end{aligned}$$

- **Categorical data:** Multinomial models usually use the logit link, but the definition of the referent group will vary.

List of Generalized Linear Models

Model	Random	Link
Linear Regression	Normal	Identity
Logistic Regression	Binomial	Logit
Multinomial response	Multinomial	Cumulative Logit
Multinomial response	Multinomial	Baseline Logit
Poisson Regression	Poisson	Log
Negative Binomial Regression	Negative Binomial	Log

Table: GLM Table based on Agresti (2002), pg. 118

Other GLM's include the geometric, zero-inflated poisson, and zero-inflated negative binomial.