BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 15

Review of Logistic and Poisson Regression Models

Introduction to Generalized Linear Models

Review: Logistic Regression

Let Y be a binary response, where

$Y = 1$ represents a "success"; $Y = 0$ represents a "failure".

Then the mean of the binary response variable, denoted π, is the *proportion* of successes or the probability that the response takes on the value 1.

That is,

$$\pi = E(Y) = \Pr(Y = 1) = \Pr(\text{"success"})$$

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

To do this, we can use *logistic regression*.

We consider a logistic regression model where

$$\ln\left[\pi/\left(1-\pi\right)\right] = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

This model accommodates the constraint that π is restricted to values between 0 and 1.

Recall that $\pi/\left(1-\pi\right)$ is defined as the odds of success.

Note that the relationship between π and the covariates is non-linear.

We can use logistic regression when:

$Y \sim$ Bernoulli(π) distribution [$Y = 0/1$]:

$$P(Y=y) = \pi^y \left(1-\pi\right)^{(1-y)}, \quad \mathrm{E}(Y) = \pi, \quad \mathrm{Var}\left(Y\right) = \pi\left(1-\pi\right)$$

$Y \sim \mathrm{Binomial}\left(n, \pi\right)$ [$Y = 0, 1, \ldots, n$ successes out of n trials]

$$P(Y=y;n) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y \left(1-\pi\right)^{(n-y)}, \quad \mathrm{E}(Y) = n\pi, \quad \mathrm{Var}\left(Y\right) = n\pi\left(1-\pi\right)$$

Under either of these assumptions we can use ML estimation to obtain estimates of the logistic regression parameters.

Given the logistic regression model

$$\ln\left[\pi/\left(1-\pi\right)\right] = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

the population intercept, β_1, has interpretation as the log odds of success when all of the covariates take on the value zero.

The population slope, say β_k, has interpretation in terms of the change in log odds of success for a single-unit change in X_k given that all of the other covariates remain constant.

When one of the covariates is dichotomous, say X_2, then β_2 has a special interpretation:

$\exp\left(\beta_2\right)$ is the *odds ratio* or ratio of odds of success for the two possible levels of X_2 (given that all of the other covariates remain constant).

Review: Poisson Regression

In Poisson regression, the response variable is a count (e.g. number of cases of a disease in a given period of time).

The Poisson distribution provides the basis of likelihood-based inference.

Often the counts may be expressed as *rates*.

That is, the count or absolute number of events is often not satisfactory because any comparison depends almost entirely on the sizes of the groups (or the "time at risk") that generated the observations.

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.

The Poisson regression model has two components:

1. 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

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

   Here, $E(Y) = \lambda, \quad \text{Var}(Y) = \lambda$

   The expected rate is given by λ/t, where t is a relevant baseline measure (e.g., t might be the number of persons or the number of person-years of observation).

2. $\ln(\lambda/t) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

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

   $$\ln(\lambda) = \ln(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

   where the 'coefficient' associated with $\ln(t)$ is fixed to be 1.

   This adjustment term is known as an "offset".

Person - Years	Smoking	Blood Pressure	Behavior	CHD
5268.2	0	0	0	20
2542.0	10	0	0	16
1140.7	20	0	0	13
614.6	30	0	0	3
4451.1	0	0	1	41
2243.5	10	0	1	24
1153.6	20	0	1	27
925.0	30	0	1	17
1366.8	0	1	0	8
497.0	10	1	0	9
238.1	20	1	0	3
146.3	30	1	0	7
1251.9	0	1	1	29
640.0	10	1	1	21
374.5	20	1	1	7
338.2	30	1	1	12

Therefore, modelling λ (or λ/t) with a log 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 expected count (or rate).

Note that the relationship between λ (or λ/t) and the covariates is non-linear.

We can use ML estimation to obtain estimates of the Poisson regression parameters, under the assumption that the responses are *Poisson* random variables.

Given the Poisson regression model

$$\ln(\lambda/t) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

the population intercept, β_1, has interpretation as the log expected rate when all the covariates take on the value zero.

The population slope, say β_k, has interpretation in terms of the change in log expected rate for a single-unit change in X_k given that all of the other covariates remain constant.

When one of the covariates is dichotomous, say X_2, then β_2 has a special interpretation:

$\exp(\beta_2)$ is the (incidence) rate ratio for the two possible levels of X_2 (given that all of the other covariates remain constant).

Overdispersion

Count data (or counts of number of successes) often have variability that far exceeds that predicted by Poisson (or binomial) distribution.

This phenomenon is referred to as *overdispersion*.

Although underdispersion can also arise, it is far less common.

Failure to account for overdispersion has negligible impact of the estimated regression coefficients.

Neglecting overdispersion results in standard errors being underestimated and potentially misleading inferences (e.g., confidence intervals that are too narrow and p-values that are too small).

Example: *Clinical Trial of Antibiotics for Leprosy*

Placebo-controlled clinical trial of 30 patients with leprosy at the Eversley Childs Sanitorium in the Philippines.

Participants were randomized to either of two antibiotics (denoted treatment drug A and B) or to a placebo (denoted treatment drug C).

Baseline data on number of leprosy bacilli at 6 sites of body were recorded.

After several months of treatment, number of bacilli were recorded a second time.

Outcome: Total count of number of leprosy bacilli at 6 sites.

Table 1: Mean count of leprosy bacilli at six sites of the body (and variance) post-treatment.

Treatment Group	Post-Treatment
Drug A (Antibiotic)	5.3
	(21.6)
Drug B (Antibiotic)	6.1
	(37.9)
Drug C (Placebo)	12.3
	(51.1)

Consider outcome (post-treatment) at end of study.

Variability is approximately 4 to 6 times larger than that predicted by Poisson variation.

Adjustments to nominal standard errors to account for overdispersion can be made either by including a scale factor ϕ in specification of the Poisson variance,

$$\text{Var}(Y_i) = \phi\,\mu_i,$$

or by basing standard errors on the so-called "sandwich" estimator of $\text{Cov}(\widehat{\beta})$.

Introduction to Generalized Linear Models

Generalized linear models are a class of regression models; they include the standard linear regression model but also many other important models:

- Linear regression for continuous data
- Logistic regression for binary data
- Loglinear/Poisson regression models for count data

Generalized linear models extend the methods of regression analysis to settings where the outcome variable can be categorical.

Later, we consider extensions of generalized linear models to longitudinal data.

Notation for Generalized Linear Models

Assume N independent observations of a single response variable, Y_i.

Associated with each response, Y_i, there is a $p \times 1$ vector of covariates, $X_{i1}, ..., X_{ip}$.

Goal: Primarily interested in relating the mean of Y_i,
$\mu_i = E(Y_i|X_{i1}, ..., X_{ip})$, to the covariates.

In generalized linear models:

(i) the distribution of the response is assumed to belong to a family of distributions known as the exponential family, e.g., normal, Bernoulli, binomial, and Poisson distributions.

(ii) A transformation of the mean response, μ_i, is then linearly related to the covariates, via an appropriate link function:

$$g(\mu_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where link function $g(\cdot)$ is a known function, e.g., $\log(\mu_i)$.

Mean and Variance of Exponential Family Distributions

Exponential family distributions share some common statistical properties.

The variance of Y_i can be expressed in terms of

$$\text{Var}\,(Y_i) = \phi\, v(\mu_i),$$

where the scale parameter $\phi > 0$.

The variance function, $v(\mu_i)$, describes how the variance of the response is functionally related to μ_i, the mean of Y_i.

Link Function

The link function applies a transformation to the mean and then links the covariates to the transformed mean,

$$g(\mu_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where link function $g(\cdot)$ is known function, e.g., $\log(\mu_i)$.

This implies that it is the transformed mean response that changes linearly with changes in the values of the covariates.

Canonical link and variance functions for the normal, Bernoulli, and Poisson distributions.

Distribution	**Var. Function,** $v(\mu)$	**Canonical Link**
Normal	$v(\mu) = 1$	Identity: $\mu = \eta$
Bernoulli	$v(\mu) = \mu(1-\mu)$	Logit: $\log\left[\frac{\mu}{(1-\mu)}\right] = \eta$
Poisson	$v(\mu) = \mu$	Log: $\log(\mu) = \eta$

where $\eta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.

Common Examples

Normal distribution:

If we assume that $g(\cdot)$ is the identity function,

$$g(\mu) = \mu$$

then

$$\mu_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

gives the standard linear regression model, with $\text{Var}(Y_i) = \phi$.

Note: Variance is unrelated to the mean.

Bernoulli distribution:

For the Bernoulli distribution, $0 < \mu_i < 1$, and so we would prefer a link function that transforms the interval $[0, 1]$ on to the entire real line $(-\infty, \infty)$:

$$\text{logit} : \ln\left[\mu_i / \left(1 - \mu_i\right)\right]$$

$$\text{probit} : \Phi^{-1}\left(\mu_i\right)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.
If we assume a logit link function then

$$\log\left[\frac{\mu_i}{(1-\mu_i)}\right] = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

yields logistic regression model, with $\text{Var}\left(Y_i\right) = \mu_i(1 - \mu_i)$ (Bernoulli variance).

Poisson distribution:

For the Poisson distribution, $\mu_i > 0$, and so we would prefer a link function that transforms the interval $(0, \infty)$ on to the entire real line $(-\infty, \infty)$.

If we assume a log link function then

$$\log\left(\mu_i\right) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

yields Poisson or loglinear regression model, with $\text{Var}\left(Y_i\right) = \mu_i$ (Poisson variance).

Summary

In generalized linear models:

(i) response assumed to have exponential family distribution, e.g., normal, Bernoulli, binomial, and Poisson distributions.

(ii) transformed mean response is linearly related to the covariates, via an appropriate link function:

$$g(\mu_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

PROC GENMOD in SAS

Table 2: Illustrative commands for logistic regression using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;
   CLASS   group;
   MODEL   y=group / DIST=BINOMIAL   LINK=LOGIT;
```

PROC GENMOD in SAS

Table 3: Illustrative commands for log-linear regression, with an offset, using PROC GENMOD in SAS.

```
PROC GENMOD;
   CLASS   group;
   MODEL   y=group / DIST=POISSON   LINK=LOG   OFFSET=logtime;
```


Extensions of Generalized Linear Models to Longitudinal Data

When the response variable is categorical (e.g., binary and count data), generalized linear models (e.g., logistic regression) can be extended to handle the correlated outcomes.

However, non-linear transformations of the mean response (e.g., logit) raise additional issues concerning the interpretation of the regression coefficients.

As we will see, different models for discrete longitudinal data have somewhat different targets of inference.

Generalized Linear Models for Longitudinal Data

Next, we focus on a number of distinct approaches for analyzing longitudinal responses.

These approaches can be considered extensions of generalized linear models to correlated data.

The main emphasis will be on discrete response data, e.g., count data or binary responses.

Note: In linear (mixed effects) models for continuous responses, the interpretation of the regression coefficients is independent of the correlation among the responses.

With discrete response data, this is no longer the case.

With non-linear models for discrete data, different approaches for accounting for the correlation lead to models having regression coefficients with distinct interpretations.

We will return to this important issue later in the course.

In the remainder of this lecture, we will briefly survey three main extensions of generalized linear models.

Suppose that $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$ is a vector of correlated responses from the i^{th} subject.

To analyze such correlated data, we must specify, or at least make assumptions about, the multivariate or joint distribution,

$$f(Y_{i1}, Y_{i2}, \ldots, Y_{in})$$

The way in which the multivariate distribution is specified yields three somewhat different analytic approaches:

1. Marginal Models

2. Mixed Effects Models

3. Transitional Models

Marginal Models

One approach is to specify the marginal distribution at each time point:

$$f(Y_{ij}) \text{ for } j = 1, 2, \ldots, n$$

along with some assumptions about the covariance structure of the observations.

The basic premise of marginal models is to make inferences about population averages.

The term "marginal" is used here to emphasize that the mean response modelled is conditional only on covariates and not on other responses (or random effects).

Illustration

Consider the *Oral Treatment of Toenail Infection* study.

Randomized, double-blind, parallel-group, multicenter study of 294 patients comparing 2 oral treatments (denoted A and B) for toenail infection.

Outcome variable: Binary variable indicating presence of onycholysis (separation of the nail plate from the nail bed).

Patients evaluated for degree of onycholysis (separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4, 8, 12, 24, 36, and 48.

Interested in the rate of decline of the proportion of patients with onycholysis over time and the effects of treatment on that rate.

Assume that the marginal probability of onycholysis follows a logistic model,

$$\text{logit}\{Pr(Y_{ij} = 1)\} = \beta_1 + \beta_2 \text{Month}_{ij} + \beta_3 \text{Trt}_i * \text{Month}_{ij}$$

where $Trt = 1$ if treatment group B and 0 otherwise.

This is an example of a marginal model.

Note, however, that the covariance structure remains to be specified.

Mixed Effects Models

Another possibility is to assume that a subset of the regression parameters in the generalized linear model vary from subject to subject.

Specifically, we could assume that the data for a single subject are independent observations from a distribution belonging to the exponential family, but that the regression coefficients can vary from person to person.

That is, conditional on the random effects, it is assumed that the responses for a single subject are independent observations from a distribution belonging to the exponential family.

Illustration

Consider the *Oral Treatment of Toenail Infection* study.

Suppose, for example, that the probability of onycholysis for participants in the study is described by a logistic model, but that the risk for an individual depends on her latent (perhaps environmentally and genetically determined) "random response level".

Then we might consider a model where

$$\text{logit}\{Pr(Y_{ij} = 1|b_i)\} = \beta_1 + \beta_2 \text{Month}_{ij} + \beta_3 \text{Trt}_i * \text{Month}_{ij} + b_i$$

Note that such a model also requires specification of the random effects distribution, $F(b_i)$.

This is an example of a generalized linear mixed effects model.

Transitional (Markov) Models

Finally, another approach is to express the joint distribution as a series of conditional distributions,

$$f\left(Y_{i1}, Y_{i2}, \ldots, Y_{in}\right) = f\left(Y_{i1}\right) f\left(Y_{i2}|Y_{i1}\right) \cdots f\left(Y_{in}|Y_{i1}, \ldots, Y_{i,n-1}\right)$$

This is known as a transitional model (or a model for the transitions) because it represents the probability distribution at each time point as conditional on the past.

This provides a complete representation of the joint distribution.

Illustration

Consider the *Oral Treatment of Toenail Infection* study.

We could write the probability model as

$$f\left(Y_{i1}|X_i\right) f\left(Y_{i2}|Y_{i1}, X_i\right) f\left(Y_{i3}|Y_{i1}, Y_{i2}, X_i\right) \cdots f\left(Y_{i7}|Y_{i1}, Y_{i2}, ..., Y_{i6}, X_i\right)$$

That is, the probability of onycholysis at time 2 is modeled conditional on presence/absence of onycholysis at time 1, and so on.

For example, a "1st-order" logistic model allowing dependence only on previous response, is given by

$$\text{logit}\{Pr(Y_{ij} = 1|Y_{i,j-1})\} = \beta_1 + \beta_2 \text{Month}_{ij} + \beta_3 \text{Trt}_i * \text{Month}_{ij} + \beta_4 Y_{i,j-1}$$

Summary

We have discussed the main features of generalized linear models

We have briefly outlined three main extensions of generalized linear models to longitudinal data:

1. Marginal Models

2. Mixed Effects Models

3. Transitional Models

In the remainder of the course we focus on (i) Marginal Models, and (ii) Mixed Effects Models.

In general, transitional models are somewhat less useful for modelling covariate effects.

Specifically, inferences from a transitional model can be potentially misleading if a treatment or exposure changes risk throughout the follow-up period.

In that case, the conditional risk, given previous history of the outcome, is altered somewhat less strikingly.