Introduction

Dipankar Bandyopadhyay, Ph.D.

Department of Biostatistics,
Virginia Commonwealth University

Course logistics

Let Y be a discrete random variable with $f(y) = P(Y = y) = p_y$.

Then, the expectation of Y is defined as

$$E(Y) = \sum_y y f(y)$$

Similarly, the Variance of Y is defined as

$$\begin{aligned} Var(Y) &= E[(Y - E(Y))^2] \\ &= E(Y^2) - [E(Y)]^2 \end{aligned}$$

Conditional probabilities

- Let A denote the event that a randomly selected individual from the "population" has heart disease.
- Then, $P(A)$ is the probability of heart disease in the "population".
- Let B denote the event that a randomly selected individual from the population has a defining characteristics such as smoking
- Then, $P(B)$ is the probability of smoking in the population
- Denote

$$P(A|B) = \begin{array}{l}\text{probability that a randomly selected individual}\\ \text{has characteristic A, given that he has characteristic B}\end{array}$$

- Then by definition,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(AB)}{P(B)}$$

provided that $P(B) \neq 0$

- $P(A|B)$ could be interpreted as the probability of that a smoker has heart disease

Associations

- The two characteristics, A and B are associated if

$$P(A|B) \neq P(A)$$

- Or, in the context of our example–the rate of heart disease depends on smoking status
- If $P(A|B) = P(A)$ then A and B are said to be independent

Bayes' theorem

- Note that
$$P(A|B) = \frac{P(AB)}{P(B)}$$
  and
$$P(B|A) = \frac{P(BA)}{P(A)}$$
- So
$$P(A|B)P(B) = P(B|A)P(A)$$
- and
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$
- which is known as Bayes' theorem

Law of Total Probability

- Suppose event B is made up of k mutually exclusive and exhaustive strata, identified by $B_1, B_2, \ldots B_k$
- If event A occurs at all, it must occur along with one (and only one) of the k exhaustive categories of B.
- Since $B_1, B_2, \ldots B_k$ are mutually exclusive

$$\begin{aligned} P(A) &= P[(A \text{ and } B_1) \text{ or } (A \text{ and } B_2) \text{ or } \ldots (A \text{ and } B_k)] \\ &= P(AB_1) + P(AB_2) + \ldots + P(AB_k) \\ &= \sum_{i=1}^{k} P(A|B_i)P(B_i) \end{aligned}$$

- which is known as the total law of probability
- A special case when $k = 2$ is

$$P(A) = P(A|B)P(B) + P(A|B')P(B')$$

where B' is read "not B" – also view this as a weighted average

Application to screening tests

- A frequent application of Bayes' theorem is in evaluating the performance of a diagnostic test used to screen for diseases
- Let D^+ be the event that a person does have the disease;
- D^- be the event that a person does NOT have the disease;
- T^+ be the event that a person has a POSITIVE test; and
- T^- be the event that a person has a NEGATIVE test
- There are 4 quantities of interest:
  1. Sensitivity
  2. Specificity
  3. Positive Predictive Value (PPV)
  4. Negative Predictive Value (NPV)

Sensitivity and Specificity

- Sensitivity is defined as the probability a test is positive given disease

$$\text{Sensitivity} = P(T^{+}|D^{+})$$

- Specificity is defined as the probability of a test being negative given the absence of disease

$$\text{Specificity} = P(T^{-}|D^{-})$$

- In Practice, you want to know disease status given a test result

PPV and NPV

- PPV is defined as the proportion of people with a positive test result that actually have the disease, which is $P(D^+|T^+)$
- By Bayes' theorem,

$$\text{PPV} = P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+)}$$

- NPV is defined as the proportion of people among those with a negative test who truly do not have the disease $(P(D^-|T^-))$
- Which by Bayes' theorem is

$$\begin{aligned}\text{NPV} &= P(D^-|T^-) \\ &= \frac{P(T^-|D^-)\cdot P(D^-)}{P(T^-)} \\ &= \frac{P(T^-|D^-)\cdot(1-P(D^+))}{1-P(T^+)}\end{aligned}$$

As a function of disease prevalence

- For both PPV and NPV, the disease prevalence ($P(D^+)$) influences the value of the screening test.
- Consider the following data

	Test result		
Disease status	Positive	Negative	Total
Present	950	50	1000
Absent	10	990	1000

- Sensitivity and Specificity for this test are

$$\text{Sen} = P(T^+|D^+) = 950/1000 = 0.95$$

and

$$\text{Spec} = P(T^-|D^-) = 990/1000 = 0.99$$

- However, the real question is what is the probability that an individual has the disease given a positive test result.

- With some easy algebra (substituting definitions into the previous equations), it can be shown that

$$\text{PPV} = \frac{Sens \cdot \Pi}{Sens \cdot \Pi + (1 - Spec)(1 - \Pi)}$$

- and

$$\text{NPV} = \frac{Spec \cdot (1 - \Pi)}{Spec \cdot (1 - \Pi) + (1 - Sens) \cdot \Pi}$$

  where Π is the disease prevalence $(P(D^+))$

- Thus, the PPV and NPV for rare to common disease could be calculated as follows:

Π	PPV	NPV
1/1,000,000	0.0001	1.0
1/500	0.16	0.99990
1/100	0.49	0.99949

Interpretation?

- For a rare disease that affects only 1 in a million,
  1. A negative test result almost guarantees the individual is free from disease (NOTE: this is a different conclusion of a 99% specificity)
  2. A positive test result still only indicates that you have a probability of 0.0001 of having the disease (still unlikely–which is why most screening tests indicate that "additional verification may be necessary")
- However, if the disease is common (say 1 in 100 have it)
  1. A negative test result would correctly classify 9995 out of 10,000 as negative, but 5 of 10,000 would be wrongly classified (i.e., they are truly positive and could go untreated)
  2. However, of 100 people that do have a positive test, only 49 would actually have the disease (51 would be wrongly screened)
- Does the test "work"
- It "depends"

Application to Pregnancy Tests

- Most home pregnancy tests claims to be “over 99% **accurate**”
- By accurate, the manufactures mean that 99% of samples are “correctly” classified (i.e., pregnant mothers have a positive test, non-pregnant mothers have a negative test)
- This measure is flawed in that it is highly dependent on the number of cases (i.e., pregnant mothers) and controls (i.e., non-pregnant mothers) – FYI: we’ll revisit this concept again in future lectures
- However, for sake of illustration, lets consider a sample of 250 pregnant mothers and 250 non-pregnant mothers

Example Data–Based on at home pregnancy tests

Suppose we have the following data observed in a clinical trial:

	Truth		
	Pregnant	Not Pregnant	
Test +	N_{++}	b	
Test -	a	N_{--}	
	250	250	500

We know that we have 99% accuracy (because the manufactures tell us so), we have a constraint

$$\frac{N_{++} + N_{--}}{500} \geq 0.99$$

so

$$N_{++} + N_{--} \geq 495$$

and for illustrative purposes, let $a = 3$ and $b = 2$ so that the following table results.

	Truth		
	Pregnant	Not Pregnant	
Test +	247	2	249
Test -	3	248	251
	250	250	500

Then

$$Sens = P(T^+|D^+) = 247/250 = 0.988$$

and

$$Spec = P(T^-|D^-) = 248/250 = 0.992$$

Using these values and simplifying the previous equations for PPV and NPV,

$$PPV = \frac{0.988\Pi}{0.980\Pi + 0.008}$$

$$NPV = \frac{0.992 - 0.992\Pi}{0.992 - 0.98\Pi}$$

where Π is again the "disease rate" (or in this case, the probability of being pregnant)

Π	PPV	NPV
0.001	0.110022	0.999988
0.01	0.555056	0.999878
0.1	0.932075	0.998658
0.5	0.991968	0.988048

- Here, the "population" at risk is those females, of childbearing age, who engaged in sexual activity during the previous menstrual cycle, and are at least 2 days late in the new cycle.
- The success rate of birth control may be in the range of 99% and unprotected sex may be in the range of (0.1-0.5)
- How do you feel about the marketing claim that the product is "over 99% accurate"?

Different Case-Control Ratio

	Truth		
	Pregnant	Not Pregnant	
Test +	397	2	399
Test -	3	98	101
	400	100	500

Then

$$Sens = P(T^+|D^+) = 297/400 = 0.9925$$

and

$$Spec = P(T^-|D^-) = 98/100 = 0.98$$

*Note: Sensitivity is now higher and specificity is lower than previously assumed

Π	PPV	NPV
0.001	0.047324	0.999992
0.01	0.333894	0.999923
0.1	0.846482	0.99915
0.5	0.980247	0.992405

Before we begin

- Lectures will be primarily from the text, and (usually) posted the night before.
- Sample SAS code is provided on Dr. Agresti's website `http://www.stat.ufl.edu/∼aa/cda/cda.html`, and in my notes. There is also a link to a large PDF file with sample `R` code.
- For a better understanding, read the lecture notes AND the text.

Categorical Data Analysis (CDA) ?

- What are categorical data?
- Agresti's answer: a variable with a measurement scale consisting of a set of categories

CDA definitions continued...

Response data considered in regression and ANOVA are continuous. Examples:

- cholesterol level (milligrams per deciliter)
- lifetime of a lab rat (in weeks)
- money spent on breakfast cereal (U.S. $)

A *categorical* variable takes on one of a (usually finite) number of categories, or levels. Examples:

- eye color (blue, brown, green, other)
- political affiliation (Democrat, Republican, other)
- cholesterol level (low, normal, high)

Note that a variable can be continuous or categorical depending on how it's defined.

Quantitative vs. Qualitative Variable Distinctions

Qualitative Variables: Distinct categories differ in quality, not in quantity

Quantitative Variables: Distinct levels have differing amounts of the characteristic of interest.

Clearly, a qualitative variable is synonymous with ”nominal” (black, white, green, blue). Also, an interval variable is clearly quantitative (weight in pounds).

However, ordinal variables are a hybrid of both a quantitative and qualitative features. For example, ‘small, medium and large’ can be viewed as a quantitative variable.

At this point, the utility in the variable descriptions may appear unnecessary. However, as the course progresses, the statistical methods presented will be appropriate for a specific classification of data.

Response or Explanatory?

Variables can be classified as a *response* or *explanatory*.

In regression models we seek to model a response as a stochastic function of explanatory variables, or *predictors*.

In this course the response will be categorical and the predictors can be categorical, continuous, or discrete.

For example, if we wanted to model political affiliation as a function of gender and annual salary, the response would be (Republican, Democrat, other), and the two predictors would be *annual salary* (essentially continuous) and the categorical *gender* (male, female).

More examples: Nominal verses Ordinal

Nominal variables have no natural ordering to them. e.g. eye color (blue, brown, other), political affiliation (Democrat, Republican, other), favorite music type (jazz, folk, rock, rap, country, bluegrass, other), gender (male, female).

Ordinal variables have an obvious order to them. e.g. cancer stage (I, II, III, IV), a taste response to a new salsa (awful, below average, average, good, delicious).

Interval variables are ordinal variables that also have a natural scale attached to them. e.g. diastolic blood pressure, number of years of post high school education. Interval variables are typically discrete numbers that comprise an interval.
Read: Sections 1.1.3, 1.1.4, 1.1.5.

Core Discrete Distributions for CDA

There are three core discrete distributions for categorical data analysis

1. Binomial (with the related Bernoulli distribution)
2. Multinomial
3. Poisson

We will explore each of these in more detail.

Bernoulli Trials

Consider the following,

- n independent patients are enrolled in a single arm (only one treatment) Phase II oncology study.
- The outcome of interest is whether or not the experimental treatment can shrink the tumor.
- Then, the outcome for patient i is

$$Y_i = \begin{cases} 1 \text{ if new treatment shrinks tumor (success)} \\ 0 \text{ if new treatment does not shrinks tumor (failure)} \end{cases},$$

$i = 1, \ldots, n$

Each Y_i is assumed to be independently, identically distributed as a Bernoulli random variables with the probability of success as

$$P(Y_i = 1) = p$$

and the probability of failure is

$$P(Y_i = 0) = 1 - p$$

Then, the probability function is Bernoulli

$$P(Y_i = y) = p^y (1-p)^{1-y} \qquad \text{for } y = 0, 1$$

and is denoted by

$$Y_i \sim Bern(p)$$

Properties of Bernoulli

- MEAN

$$\begin{aligned} E(Y_i) &= 0 \cdot P(Y_i = 0) + 1 \cdot P(Y_i = 1) \\ &= 0(1-p) + 1p \\ &= p \end{aligned}$$

- VARIANCE

$$\begin{aligned} Var(Y_i) &= E(Y_i^2) - [E(Y_i)]^2 \\ &= E(Y_i) - [E(Y_i)]^2 \ \ ; \ \text{since } Y_i^2 = Y_i \\ &= E(Y_i)[1 - E(Y_i)] \\ &= p(1-p) \end{aligned}$$

Binomial Distribution

Let Y be defined as

$$Y = \sum_{i=1}^{n} Y_i,$$

where n is the number of bernoulli trials. We will use Y (the number of successes) to form test statistics and confidence intervals for p, the probability of success.

Example 2,
Suppose you take a sample of n independent biostatistics professors to determine how many of them are nerds (or geeks).

We want to estimate the probability of being a nerd given you are a biostatistics professor.

What is the distribution of the number of successes,

$$Y = \sum_{i=1}^{n} Y_i,$$

resulting from n identically distributed, independent trials with

$$Y_i = \begin{cases} 1 \text{ if professor } i \text{ is a nerd (success)} \\ 0 \text{ if professor } i \text{ is not a nerd (failure)} \end{cases} .$$

and

$$p = P(Y_i = 1); \qquad (1 - p) = P(Y_i = 0)$$

for all $i = 1, \ldots, n$

The probability function can be shown to be binomial:

$$P(Y=y)=\left(\begin{array}{c} n \\ y \end{array}\right) p^y(1-p)^{n-y}=\frac{n!}{y!(n-y)!}p^y(1-p)^{n-y},$$

where

$$y=0,1,2,\ldots,n$$

and
the number

$$\left(\begin{array}{c} n \\ y \end{array}\right)=\frac{n!}{(n-y)!y!}$$

is the number of ways of partitioning n objects into two groups; one group of size y, and the other of size $(n-y)$.
The distribution is denoted by

$$Y \sim Bin(n,p)$$

Properties of the Binomial

- MEAN

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^{n} Y_i\right) \\ &= \sum_{i=1}^{n} E(Y_i) \\ &= \sum_{i=1}^{n} p = np \end{aligned}$$

  [Recall, the expectation of a sum is the sum of the expectations]

- VARIANCE

$$\begin{aligned} Var(Y) &= Var\left(\sum_{i=1}^{n} Y_i\right) \\ &= \sum_{i=1}^{n} Var(Y_i) \\ &= \sum_{i=1}^{n} p(1-p) = np(1-p) \end{aligned}$$

  [Variance of a sum is the sum of the variances if observations are independent)

Multinomial

Often, a categorical may have more than one outcome of interest. Recall the previous oncology trial where Y_i was defined as

$$Y_i = \begin{cases} 1 \text{ if new treatment shrinks tumor (success)} \\ 0 \text{ if new treatment does not shrinks tumor (failure)} \end{cases}$$

However, sometimes is may be more beneficial to describe the outcome in terms of

$$Y_i = \begin{cases} 1 \text{ Tumor progresses in size} \\ 2 \text{ Tumor remains as is} \\ 3 \text{ Tumor decreases in size} \end{cases}$$

Multinomial

Let $y_{ij} = 1$ if subject i has outcome j and $y_{ij} = 0$ else. Then

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ic})$$

represents a multinomial trial, with $\sum_j y_{ij} = 1$ and c representing the number of potential levels of Y.

Let $n_j = \sum_i y_{ij}$ denote the number of trials having outcome in category j. The counts $(n_1, n_2, \ldots, n_c)$ have the multinomial distribution.

$$P(n_1, n_2, \cdots, n_{c-1}) = \left(\frac{n!}{n_1! n_2! \cdots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

How many free parameters among the π's?

Special Case of a Multinomial

When $c = 2$, then

$$P(n_1) = \left(\frac{n!}{n_1!n_2!}\right)\pi_1^{n_1}\pi_2^{n_2}$$

Due to the constraints $\sum_c n_c = n$ and $\sum_c \pi = 1$, $n_2 = n - n_1$ and $\pi_2 = 1 - \pi_1$.

Therefore,

$$P(n_1) = \left(\frac{n!}{n_1!(n - n_1!)}\right)\pi_1^{n_1}(1 - \pi_1)^{n-n_1}$$

Note: For most of the class, I will use p for probability, Agresti tends to use π

Poisson

Sometimes, count data does not arrive from a fixed number of trials. For example,
Let Y = number of babies born at VCU in a given week.

Y does not have a predefined maximum and a key feature of the Poisson distribution is that the variance equals its mean.

The probability that $Y = 0, 1, 2, \ldots$ is written as

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

where $\mu = E(Y) = Var(Y)$.

Proof of Expectation

$$\begin{aligned}
E[Y] &= \sum_{i=0}^{\infty} \frac{ie^{-\mu}\mu^i}{i!} \\
&= \frac{0 \cdot e^{-\mu}}{0!} + \sum_{i=1}^{\infty} \frac{ie^{-\mu}\mu^i}{i!} \qquad \text{[See Note 1]} \\
&= 0 + \mu e^{-\mu} \sum_{i=1}^{\infty} \frac{\mu^{i-1}}{(i-1)!} \\
&= \mu e^{-\mu} \sum_{j=0}^{\infty} \frac{\mu^j}{j!} \qquad \text{[See Note 2]} \\
&= \mu \quad \text{[See Note 3]}
\end{aligned}$$

Notes:

1. $0! = 1$ and we separated the 1^{st} term (i=0) of the summation out
2. Let $j = i - 1$, then if $i = 1, \ldots, \infty$, $j = 0, \ldots, \infty$
3. Since $\sum_{j=0}^{\infty} \frac{\mu^j}{j!} = e^{\mu}$ by McLaurin expansion of e^x

Overdispersion

Overdispersion: Often the variability associated with Poisson and binomial models is smaller than what is observed in real data.

The increased variance can be attributed to unmeasured, or perhaps latent regressors in the model and thus the resulting count distribution is more correctly a *mixture* of binomial or Poisson distributions, with mixing weights being the proportion of outcomes resulting from specific (unaccounted for) covariate combinations.

We will discuss testing for overdispersion in specific models and remedies later on.

Connection between Multinomial and Poisson

Let $\mathbf{Y} = (Y_1, Y_2, Y_3)$ be independent Poisson with parameters (μ_1, μ_2, μ_3).

e.g. Y_1 is number of people that fly to France from Britain this year, Y_2 the number who go by train, and Y_3 the number who take a ferry. The total number of traveling $n = Y_1 + Y_2 + Y_3$ is Pois$(\mu_1 + \mu_2 + \mu_3)$.

Conditional on n, the distribution of (Y_1, Y_2, Y_3) is multinomial with parameters n and $\boldsymbol{\pi} = (\mu_1, \mu_2, \mu_3)/\mu_+$ where $\mu_+ = \mu_1 + \mu_2 + \mu_3$.

This is especially useful in log-linear models, covered in Chapter 9.

Maximum likelihood [ML] estimation

Let the parameter vector for a model be $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ where p is the number of parameters in the model. Let the outcome variables be random variables denoted $\mathbf{y} = (y_1, \ldots, y_n)$ and the probability model denoted

$$p(y_1, \ldots, y_n | \boldsymbol{\beta}) = p(\mathbf{y} | \boldsymbol{\beta}).$$

The likelihood of $\boldsymbol{\beta}$, denoted $\mathcal{L}(\boldsymbol{\beta})$, is $\mathcal{L}(\boldsymbol{\beta}) = p(\mathbf{y}|\boldsymbol{\beta})$ thinking of data $\mathbf{y}$ as fixed.

MLE, cont.

For example, if $\mathbf{n} = (n_1, \ldots, n_c)$ is $\text{mult}(n, \boldsymbol{\pi})$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$, then $\boldsymbol{\beta} = (\pi_1, \pi_2, \ldots, \pi_{c-1})$ because there are $c - 1$ free parameters in $\boldsymbol{\pi}$.

The likelihood of $\boldsymbol{\beta}$ is simply the probability of seeing the response data given $\boldsymbol{\beta}$:

$$\mathcal{L}(\boldsymbol{\beta}) = p(n_1, \ldots, n_c | \boldsymbol{\beta}) = \left(\begin{array}{c} n \\ n_1 \cdots n_c \end{array} \right) \prod_{j=1}^{c} \pi_j^{n_j}.$$

MLE, cont.

The maximum likelihood estimator is that value of $\boldsymbol{\beta}$ that maximizes $\mathcal{L}(\boldsymbol{\beta})$ for given data:

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta} \in \mathbf{B}} \mathcal{L}(\boldsymbol{\beta}),$$

where $\mathbf{B}$ is the set of values $\boldsymbol{\beta}$ can take on.

The MLE $\hat{\boldsymbol{\beta}}$ makes the observed data as *likely as possible*. The estimator turns into an estimate when data are actually seen. For example, if $c = 3$ and $n_1 = 3$, $n_2 = 5$, $n_3 = 2$, then $\hat{\boldsymbol{\beta}} = (\hat{\pi}_1, \hat{\pi}_2) = (0.3, 0.5)$ and of course $\hat{\pi}_3 = 1 - (\hat{\pi}_1 + \hat{\pi}_2) = 0.2$. Then $p(3, 5, 2|\pi_1 = 0.2, \pi_2 = 0.5) \geq p(3, 5, 2|\pi_1 = p_1, \pi_2 = p_2)$ for all values of p_1 and p_2.

An estimator is random (i.e. before data are collected and seen they are random, and so then is any function of data) whereas an estimate is a fixed, known vector (like (0.3,0.5)).

MLE, cont.

MLEs have nice properties for most (but not all) models (p. 9):

- They have large sample normal distributions:

$$\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_p(\boldsymbol{\beta}, \text{cov}(\hat{\boldsymbol{\beta}})) \text{ where } \text{cov}(\hat{\boldsymbol{\beta}}) = \left[-E\left(\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \right) \right]_{p \times p}^{-1}.$$

- They are asymptotically consistent: $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ (in probability) as the sample size $n \to \infty$.
- They are asymptotically efficient: $\text{var}(\hat{\beta}_j)$ is smaller than the corresponding variance of other (asymptotically) unbiased estimators.

Example: MLE for Poisson data

Let $Y_i \sim \text{Pois}(\lambda t_i)$ where λ is the unknown event rate and t_i are known exposure times. Assume the $Y_1, \ldots, Y_n$ are independent.

The likelihood of λ is

$$\begin{aligned}
\mathcal{L}(\lambda) &= p(y_1, \ldots, y_n|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} e^{-t_i\lambda}(t_i\lambda)^{y_i}/y_i! \\
&= \left[\prod_{i=1}^{n} \frac{t_i^{y_i}}{y_i!}\right] e^{-\lambda \sum_{i=1}^{n} t_i} \lambda^{\sum_{i=1}^{n} y_i} = g(\mathbf{t}, \mathbf{y}) e^{-\lambda \sum_{i=1}^{n} t_i} \lambda^{\sum_{i=1}^{n} y_i}.
\end{aligned}$$

Then the log-likelihood is

$$L(\lambda) = \log g(\mathbf{t}, \mathbf{y}) - \lambda \sum_{i=1}^{n} t_i + \log(\lambda) \sum_{i=1}^{n} y_i.$$

Poisson MLE, cont.

Taking the derivative w.r.t. λ we get

$$L'(\lambda) = \frac{\partial L(\lambda)}{\partial \lambda} = -\sum_{i=1}^{n} t_i + \frac{1}{\lambda}\sum_{i=1}^{n} y_i.$$

Setting this equal to zero, plugging in $\mathbf{Y}$ for $\mathbf{y}$, and solving for λ yields the MLE

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} t_i}.$$

Now

$$\frac{\partial^2 L(\lambda)}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} y_i}{\lambda^2}.$$

Since $\sum_{i=1}^{n} Y_i \sim \text{Pois}(\lambda \sum_{i=1}^{n} t_i)$, we have

$$-E\left(\frac{\partial^2 L(\lambda)}{\partial \lambda^2}\right) = E\left(\frac{\sum_{i=1}^{n} Y_i}{\lambda^2}\right) = \frac{\lambda \sum_{i=1}^{n} t_i}{\lambda^2} = \frac{\sum_{i=1}^{n} t_i}{\lambda}.$$

Poisson MLE, cont.

The variance of $\hat{\lambda}$ is given by the 'inverse' of this 'matrix'

$$\text{var}(\hat{\lambda}) \overset{\bullet}{=} \frac{\lambda}{\sum_{i=1}^{n} t_i}.$$

The large sample normal result tells us

$$\hat{\lambda} \overset{\bullet}{\sim} N\left(\lambda, \frac{\lambda}{\sum_{i=1}^{n} t_i}\right).$$

The standard deviation of $\hat{\lambda}$ is estimated to be $\text{sd}(\hat{\lambda}) = \sqrt{\frac{\lambda}{\sum_{i=1}^{n} t_i}}$. Since we do not know λ, the standard deviation is estimated by the *standard error* obtained from estimating λ by its MLE:

$$\text{se}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{\sum_{i=1}^{n} t_i}} = \sqrt{\frac{\sum_{i=1}^{n} y_i}{[\sum_{i=1}^{n} t_i]^2}} = \frac{\sqrt{\sum_{i=1}^{n} y_i}}{\sum_{i=1}^{n} t_i}.$$

Poisson MLE, cont.

Example: Say that we record the number of adverse medical events (e.g. operating on the wrong leg) from a hospital over $n = 3$ different times: $t_1 = 1$ week in 2013, $t_2 = 4$ weeks in 2014, and $t_3 = 3$ weeks in 2015. We'll assume that the adverse surgical event rate λ (events/week) does not change over time and that event counts in different time periods are independent.

Then $Y_i \sim \text{Pois}(t_i\lambda)$ for $i = 1, 2, 3$. Say we observe $y_1 = 0$, $y_2 = 3$, and $y_3 = 1$. Then $\hat{\lambda} = (0 + 3 + 1)/(1 + 4 + 3) = 4/8 = 0.5$ event/week, or one event every other week. Also, $\text{se}(\hat{\lambda}) = \sqrt{4}/8 = 0.25$.

The large sample result tells us then (before data are collected and $\mathbf{Y} = (Y_1, Y_2, Y_3)$ is random) that

$$\hat{\lambda} \overset{\bullet}{\sim} N(\lambda, 0.25^2),$$

useful for constructing hypothesis tests and confidence intervals.

Wald, Likelihood Ratio, and Score tests

These are three ways to perform large sample hypothesis tests based on the model likelihood.

Wald test

Let $\mathbf{M}$ be a $m \times p$ matrix. Many hypotheses can be written $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ where $\mathbf{b}$ is a known $m \times 1$ vector.

For example, let $p = 3$ so $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. The test of $H_0 : \beta_2 = 0$ is written in matrix terms with $\mathbf{M} = (0, 1, 0)$ and $b = 0$. The hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3$ has $\mathbf{M} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Wald test, cont.

The large sample result for MLEs is

$$\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_p(\boldsymbol{\beta}, \mathrm{cov}(\hat{\boldsymbol{\beta}})).$$

So then

$$\mathbf{M}\hat{\boldsymbol{\beta}} \overset{\bullet}{\sim} N_m(\mathbf{M}\boldsymbol{\beta}, \mathbf{M}\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{M}').$$

If $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ is true then

$$\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b} \overset{\bullet}{\sim} N_m(\mathbf{0}, \mathbf{M}\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{M}').$$

So

$$W = (\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b})'[\mathbf{M}\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}})\mathbf{M}']^{-1}(\mathbf{M}\hat{\boldsymbol{\beta}} - \mathbf{b}) \overset{\bullet}{\sim} \chi^2_m.$$

W is called the Wald statistic and large values of W indicate $\mathbf{M}\boldsymbol{\beta}$ is far away from $\mathbf{b}$, i.e. that H_0 is false. The p-value for $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ is given by p-value $= P(\chi^2_m > W)$.

The simplest, most-used Wald test is the familiar test that a regression effect is equal to zero, common to multiple, logistic, Poisson, and ordinal regression models.

Score test

In general, the $\text{cov}(\hat{\beta})$ is a function of the unknown $\boldsymbol{\beta}$. The Wald test replaces $\boldsymbol{\beta}$ by its MLE $\hat{\beta}$ yielding $\widehat{\text{cov}}(\hat{\beta})$. The score test replaces $\boldsymbol{\beta}$ by the the MLE $\hat{\boldsymbol{\beta}}_0$ obtained under the constraint imposed by H_0

$$\hat{\boldsymbol{\beta}}_0 = \text{argmax}_{\boldsymbol{\beta}\in\mathbf{B}:\mathbf{M}\boldsymbol{\beta}=\mathbf{b}}\mathcal{L}(\boldsymbol{\beta}).$$

Let $\text{cov}(\boldsymbol{\beta})$ be the asymptotic covariance for *unconstrained* MLE.

The resulting test statistic

$$S = [\tfrac{\partial}{\partial\boldsymbol{\beta}}\log\mathcal{L}(\hat{\boldsymbol{\beta}}_0)]'[\text{cov}(\hat{\boldsymbol{\beta}}_0)][\tfrac{\partial}{\partial\boldsymbol{\beta}}\log\mathcal{L}(\hat{\boldsymbol{\beta}}_0)] \overset{\bullet}{\sim} \chi^2_m.$$

Sometimes it is easier to fit the reduced model rather than the full model; the score test allows testing whether new parameters are necessary from a fit of a smaller model.

Likelihood Ratio tests

The Likelihood Ratio test [LRT] is easily constructed and carried out for nested models. The full model has parameter vector $\boldsymbol{\beta}$ and the reduced model obtains when $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ holds. A common example is when $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and we wish to test $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ (e.g. a subset of regression effects are zero). Let $\hat{\boldsymbol{\beta}}$ be the MLE under the full model

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta} \in \mathbf{B}} \mathcal{L}(\boldsymbol{\beta}),$$

and $\hat{\boldsymbol{\beta}}_0$ be the MLE under the constraint imposed by H_0

$$\hat{\boldsymbol{\beta}}_0 = \text{argmax}_{\boldsymbol{\beta} \in \mathbf{B} : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}} \mathcal{L}(\boldsymbol{\beta}).$$

LRT, cont.

If $H_0 : \mathbf{M}\beta = \mathbf{b}$ is true,

$$L = -2[\log \mathcal{L}(\hat{\boldsymbol{\beta}}_0) - \log \mathcal{L}(\hat{\boldsymbol{\beta}})] \overset{\bullet}{\sim} \chi^2_m.$$

The statistic L is the likelihood ratio test statistic for the hypothesis $H_0 : \mathbf{M}\beta = \mathbf{b}$. The smallest L can be is zero when $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}$. The more different $\hat{\boldsymbol{\beta}}$ is from $\hat{\boldsymbol{\beta}}_0$, the larger L is and the more evidence there is that H_0 is false. The p-value for testing H_0 is given by $p-$value $= P(\chi^2_m > L)$.

To test whether additional parameters are necessary, LRT tests are carried out by fitting two models: a ‘full’ model with all effects and a ‘reduced’ model. In this case the dimension m of $\mathbf{M}$ is the difference in the numbers of parameters in the two models.

LRT, cont.

For example, say we are fitting the standard regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

where $e_i \overset{iid}{\sim} N(0, \sigma^2)$. Then $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$ and we want to test $\boldsymbol{\beta}_1 = (\beta_2, \beta_3) = (0, 0)$, that the 2^{nd} and 3^{rd} predictors aren't needed. This test can be written using matrices as

$$H_0 : \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

LRT, cont.

The likelihood ratio test fits the full model above and computes $L_f = \log \mathcal{L}_f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma})$.

Then the reduced model $Y_i = \beta_0 + \beta_1 x_{i1} + e_i$ is fit and $L_r = \log \mathcal{L}_r(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$ computed.

The test statistic is $L = -2(L_r - L_f)$; a p-value is computed as $P(\chi^2_2 > L)$. If the p-value is less than, say, $\alpha = 0.05$ we reject $H_0 : \beta_2 = \beta_3 = 0$.

Of course we wouldn't use this approximate LRT test here! We have outlined an approximate test, but there is well-developed theory that instead uses a different test statistic with an exact F-distribution.

Comments

Note that:

- The Wald test requires maximizing the unrestricted likelihood, and uses non-null standard error.
- The score test requires maximizing the restricted likelihood (under a nested submodel), and uses the null standard error.
- The Likelihood ratio test combines information from both [restricted and unrestricted] likelihoods.

So the likelihood ratio test uses more information and both Wald and Score tests can be viewed as approximations to the LRT.

However, SAS can "automatically" perform Wald tests of the form $H_0 : \mathbf{M}\boldsymbol{\beta} = \mathbf{b}$ in a `contrast` statement and so I often use Wald tests because they're easy to get. In large samples the tests are equivalent.

Confidence intervals

A plausible range of values for a parameter β_j (from $\boldsymbol{\beta}$) is given by a confidence interval (CI). Recall that a CI has a certain fixed probability of containing the unknown β_j before data are collected. After data are collected, nothing is random any more, and instead of "probability" we refer to "confidence."

A common way of obtaining confidence intervals is by *inverting* hypothesis tests of $H_0 : \beta_k = b$. Without delving into why this works, a $(1 - \alpha)100\%$ CI is given by those b such that the p-value for testing $H_0 : \beta_k = b$ is larger than α.

CIs, cont.

For Wald tests of $H_0 : \beta_k = b$, the test statistic is $W = (\hat{\beta}_k - b)/\text{se}(\hat{\beta}_k)$. This statistic is approximately $N(0,1)$ when $H_0 : \beta_k = b$ is true and the p-value is larger than $1 - \alpha$ only when $|W| < z_{\alpha/2}$ where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of a $N(0,1)$ random variable. This yields the well known CI

$$(\hat{\beta}_k - z_{\alpha/2}\text{se}(\hat{\beta}_k),\ \hat{\beta}_k + z_{\alpha/2}\text{se}(\hat{\beta}_k)).$$

The likelihood ratio CI operates in the same way, but the log-likelihood must be computed for all values of b. We'll explore the differences between inverting Wald, Score, and LRT for binomial data in the remainder of Chapter 1.