

STAT 426

1.3 Statistical Inference for Categorical Data (Part I)

Maximum likelihood estimation

We will mostly discuss **maximum likelihood estimation**. Assuming certain regularity conditions, the properties of the maximum likelihood estimators are:

- Large-sample normal distributions
- Asymptotically consistent (converge to the population value)
- Asymptotically efficient (lower variance than other estimators)

Maximum likelihood estimation

For our purposes, $L(\beta)$ will be well-defined and at least twice continuously differentiable.

- **Maximum Likelihood (ML) estimate:** parameter value that maximizes the likelihood function. If $\hat{\beta}$ maximizes the likelihood function $\ell(\beta)$, $\hat{\beta}$ also maximizes the logarithm of the likelihood function.
- The maximum likelihood estimate is the solution of $\partial\ell(\beta)/\partial\beta = 0$.
- If β is multidimensional, we denote the parameter vector as β and get $\hat{\beta}$ as the solution of a set of equations.

Maximum likelihood estimation

Let β a generic unknown parameter and $\hat{\beta}$ the parameter estimate:

- **Likelihood function**: the probability of observing a sample, as a function of the unknown parameter.

$\ell(\beta)$ = joint density of data at its observed values, as a function of β

log-likelihood $L(\beta) = \log(\ell(\beta))$

- The **kernel** of $\ell(\beta)$ includes only the factors that depend on β .
- Inference will involve only the kernel, so $L(\beta)$ need only be specified up to an additive constant.

For our purposes, $L(\beta)$ will be well-defined and at least twice continuously differentiable.

A **maximum likelihood estimate (MLE)** $\hat{\beta}$ maximizes $\ell(\beta)$.

$\hat{\beta}$ is usually the (unique) solution of $L'(\beta) = 0$.

Note: An MLE also maximizes the kernel.

Covariance of the ML estimators

Let $cov(\hat{\beta})$ the covariance matrix of $\hat{\beta}$.

Under some regularity conditions covariance matrix is the **inverse of the information matrix**. The (j,k) element of the information matrix can be estimated as:

$$v(\hat{\beta})_{jk} = -E\left(\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k}\right)$$

The standard errors (SE) of $\hat{\beta}$, are the square roots of the elements in the diagonal of the covariance matrix. The greater the curvature of the log likelihood, the smaller the standard errors.

Exercise: Find the likelihood function and ML estimate of the Binomial and Poisson parameter.

The **score function** is

$$u(\beta) = \frac{\partial L(\beta)}{\partial \beta}$$

The **(Fisher) information** is

$$i(\beta) = -E\left(\frac{\partial^2 L(\beta)}{\partial \beta^2}\right)$$

where the expectation is over the assumed distribution for the data when the parameter value is β .

Note: These can be found even when $L(\beta)$ is known only up to an additive constant.

If the data are from a sample of size n , we consider asymptotic behavior as $n \rightarrow \infty \dots$

Typically,

$$(\imath(\beta))^{-1} = \text{asymptotic variance of MLE } \hat{\beta}$$

in the sense that using it to “standardize” $\hat{\beta}$ results in an asymptotic limit (often normal) with variance 1. Also,

$$\sigma(\hat{\beta}) = \sqrt{(\imath(\beta))^{-1}} = \text{asymptotic standard error}$$

Can also show

$$E(u(\beta)) = 0 \qquad \text{var}(u(\beta)) = \iota(\beta)$$

where the expectations are over the assumed distribution for the data when the parameter value is β .

When the parameter value is β , $u(\beta)$ is often asymptotically normal (after appropriate standardization).

Example (Binomial Probability)

$$Y \sim \text{binomial}(n, \pi) \quad 0 < \pi < 1$$

n known π unknown

Can take

$$L(\pi) = \ln(\pi^y (1 - \pi)^{n-y}) = y \ln \pi + (n - y) \ln(1 - \pi)$$

Example (Binomial Probability)

$$Y \sim \text{binomial}(n, \pi) \quad 0 < \pi < 1$$

n known π unknown

Can take

$$L(\pi) = \ln(\pi^y (1 - \pi)^{n-y}) = y \ln \pi + (n - y) \ln(1 - \pi)$$

so that

$$u(\pi) = \frac{\partial L}{\partial \pi} = \frac{y}{\pi} - \frac{n - y}{1 - \pi} = \frac{y - n\pi}{\pi(1 - \pi)}$$

Note $E(u(\pi)) = 0$.

Example (continued)

Solving $u(\pi) = 0$ gives MLE

$$\hat{\pi} = \frac{y}{n} = \text{proportion of "successes"}$$

whenever $0 < y < n$.

(We will also formally allow $y = 0$ and $y = n$, even though $\hat{\pi} = 0$ and $\hat{\pi} = 1$ are outside the parameter space.)

Example (continued)

The information is

$$\begin{aligned} i(\pi) &= -E\left(\frac{\partial^2 L}{\partial \pi^2}\right) = E\left(\frac{Y}{\pi^2} + \frac{n - Y}{(1 - \pi)^2}\right) \\ &= \frac{n\pi}{\pi^2} + \frac{n(1 - \pi)}{(1 - \pi)^2} \\ &= \frac{n}{\pi} + \frac{n}{(1 - \pi)} = \frac{n}{\pi(1 - \pi)} \end{aligned}$$

Example (continued)

$$E(\hat{\pi}) = \pi$$

$$\begin{aligned}\text{var}(\hat{\pi}) &= \text{var}(Y/n) = n\pi(1 - \pi)/n^2 \\ &= \pi(1 - \pi)/n = (\imath(\pi))^{-1}\end{aligned}$$

so the variance is exactly the inverse information, in this case, though in general that is only approximately true.

Example (continued)

$$E(\hat{\pi}) = \pi$$

$$\begin{aligned}\text{var}(\hat{\pi}) &= \text{var}(Y/n) = n\pi(1 - \pi)/n^2 \\ &= \pi(1 - \pi)/n = (\iota(\pi))^{-1}\end{aligned}$$

so the variance is exactly the inverse information, in this case, though in general that is only approximately true. We write

$$\sigma(\hat{\pi}) = \sqrt{\pi(1 - \pi)/n}$$

By the LLN, $\hat{\pi}$ is consistent.

By the CLT, $\hat{\pi}$ is asymptotically normal.

Example (continued)

Note also that

$$\begin{aligned}\text{var}(u(\pi)) &= \text{var}\left(\frac{Y - n\pi}{\pi(1 - \pi)}\right) \\ &= \frac{1}{\pi^2(1 - \pi)^2} n\pi(1 - \pi) \\ &= \frac{n}{\pi(1 - \pi)} = i(\pi)\end{aligned}$$

So $\text{var}(u(\pi)) = i(\pi)$ exactly, as is generally true.

Likelihood Inference

Back to the general model with parameter β ...

How can we test

$$H_0 : \beta = \beta_0 \qquad H_a : \beta \neq \beta_0$$

or form a confidence interval (CI) for β ?

Likelihood Inference

Back to the general model with parameter β ...

How can we test

$$H_0 : \beta = \beta_0 \qquad H_a : \beta \neq \beta_0$$

or form a confidence interval (CI) for β ?

Three main likelihood approaches:

- Wald
- Likelihood Ratio
- Score

Wald test

These tests use the **asymptotic normality** of the maximum likelihood estimators. We want to test the null hypothesis $H_0 : \beta = \beta_0$. The test statistic:

$$z = \frac{\hat{\beta} - \beta_0}{SE}$$

for a non-zero SE, has an approximate normal distribution when $\beta = \beta_0$.

- We can obtain one-sided or two-sided P-values from the standard normal distribution function.
- For the two-sided test, the statistics z^2 has a chi-squared distribution with 1 degree of freedom.
- This type of statistic is called the **Wald** statistic.

Multivariate extension of the Wald test

We want to test $H_0 : \beta = \beta_0$.

The Wald statistic can be written as:

$$W = (\hat{\beta} - \beta_0)' [\text{cov}(\hat{\beta})]^{-1} (\hat{\beta} - \beta_0)$$

The asymptotic normal distribution for $\hat{\beta}$ implies the asymptotic chi-square distribution for W , with degrees of freedom $\text{rank}(\text{cov}(\hat{\beta}))$.

Wald

The **Wald statistic**:

$$z_W = \frac{\hat{\beta} - \beta_0}{SE} \qquad SE = \frac{1}{\sqrt{i(\hat{\beta})}}$$

(Note that SE uses $\hat{\beta}$, not β_0 .)

Wald

The **Wald statistic**:

$$z_W = \frac{\hat{\beta} - \beta_0}{SE} \qquad SE = \frac{1}{\sqrt{i(\hat{\beta})}}$$

(Note that SE uses $\hat{\beta}$, not β_0 .)

Usually

$$z_W \xrightarrow[n \rightarrow \infty]{d} N(0, 1) \quad \text{under } H_0 : \beta = \beta_0$$

so reject if $|z_W| \geq z_{\alpha/2}$ for a two-sided level α test.

The Wald test also has a chi-squared form, using

$$z_W^2 = \frac{(\hat{\beta} - \beta_0)^2}{1/\iota(\hat{\beta})} \underset{H_0}{\sim} \chi_1^2$$

Likelihood Ratio

Let

$$\Lambda = \ell(\beta_0)/\ell(\hat{\beta})$$

where $\ell(\beta_0)$ is the maximized value of the likelihood under H_0 and $\ell(\hat{\beta})$ is the maximized value over all parameter space. The ratio Λ cannot exceed 1.

The **likelihood-ratio test (LRT) chi-squared statistic**:

$$-2 \ln \Lambda = -2 \ln(\ell(\beta_0)/\ell(\hat{\beta})) = -2(L(\beta_0) - L(\hat{\beta}))$$

Likelihood Ratio

Let

$$\Lambda = \ell(\beta_0)/\ell(\hat{\beta})$$

where $\ell(\beta_0)$ is the maximized value of the likelihood under H_0 and $\ell(\hat{\beta})$ is the maximized value over all parameter space. The ratio Λ cannot exceed 1.

The **likelihood-ratio test (LRT) chi-squared statistic**:

$$-2 \ln \Lambda = -2 \ln(\ell(\beta_0)/\ell(\hat{\beta})) = -2(L(\beta_0) - L(\hat{\beta}))$$

It has an approximate χ_1^2 distribution under $H_0 : \beta = \beta_0$, and otherwise tends to be larger.

Thus, reject H_0 if

$$-2 \ln \Lambda \geq \chi_1^2(\alpha)$$

Score

The **score statistic**:

$$z_S = \frac{u(\beta_0)}{\sqrt{i(\beta_0)}}$$

(This is the score standardized under H_0 .)

Score

The **score statistic**:

$$z_S = \frac{u(\beta_0)}{\sqrt{i(\beta_0)}}$$

(This is the score standardized under H_0 .)

Under $H_0 : \beta = \beta_0$, its distribution is approximately $N(0, 1)$.
Otherwise, it tends to be further from zero.

Thus, reject H_0 if $|z_S| \geq z_{\alpha/2}$.

Score

The **score statistic**:

$$z_S = \frac{u(\beta_0)}{\sqrt{i(\beta_0)}}$$

(This is the score standardized under H_0 .)

Under $H_0 : \beta = \beta_0$, its distribution is approximately $N(0, 1)$.
Otherwise, it tends to be further from zero.

Thus, reject H_0 if $|z_S| \geq z_{\alpha/2}$.

There is also a chi-squared form:

$$z_S^2 = \frac{u(\beta_0)^2}{i(\beta_0)} \underset{H_0}{\rightsquigarrow} \chi_1^2$$

All three kinds tend to be “asymptotically equivalent” as $n \rightarrow \infty$.

For smaller n , the likelihood-ratio and score methods are preferred.