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:

$\hat{\beta}_{MLE} \sim N(\beta, Var(\hat{\beta}))$

- Large-sample normal distributions

unbiased: $\hat{\beta} \to \beta$.

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

Maximum likelihood estimation

$\hat{\beta}_{MLE} \sim N(\beta, Var(\hat{\beta}))$ → inverse of information matrix

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 $\boldsymbol{\beta}$ and get $\hat{\boldsymbol{\beta}}$ 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.

  $= \prod_{i=1}^{n} f(x_i).$

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

  log-likelihood $L(\beta) = \log(\ell(\beta))$ $= \sum_{i=1}^{n} \log(f(x_i)).$
- 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 $L'(\hat{\beta}) = 0$.

Note: An MLE also maximizes the kernel.

Covariance of the ML estimators

$\hat{\beta} \sim N((\mu, \sigma^2)^T, cov(\hat{\beta}))$

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

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:

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

The standard errors (SE) of $\hat{\boldsymbol{\beta}}$, 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

$$u(\beta) = \boxed{\frac{\partial L(\beta)}{\partial \beta} = 0}$$

Solve $\hat{\beta}_{MLE}$.

The **(Fisher) information** is

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

to derive $\mathrm{cov}(\hat{\beta})$.

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 $n \to \infty$ ...

Typically,

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

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

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

Can also show

$$E\big(u(\beta)\big) \;=\; 0 \qquad\qquad \mathrm{var}\big(u(\beta)\big) \;=\; \imath(\beta)$$

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)

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

n known $\quad$ π unknown

Can take

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

so that

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

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

Example (continued)

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

$$\hat{\pi} \quad = \quad \frac{y}{n} \quad = \quad \text{proportion of ``successes''}$$

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 $Var(\hat{\pi}) = (i(\pi))^{-1}.$

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

$$i(\pi) = Var(u(\pi)) = \frac{Var(y)}{\pi^2(1-\pi)^2} = \frac{n\pi(1-\pi)}{\pi^2(1-\pi)^2} = \frac{n}{\pi(1-\pi)}$$

Example (continued)

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

. unbiased.

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

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. $P(|\hat{\pi}-\pi|>\varepsilon) \xrightarrow{n\to\infty} 0$. $\hat{\pi} \to \pi$.

By the CLT, $\hat{\pi}$ is asymptotically normal. $\hat{\pi} \sim N(\pi, (\imath(\pi))^{-1})$.

Likelihood Inference

Back to the general model with parameter β ...

How can we test

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

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:

$$z = \frac{\hat{\beta} - \beta_0}{\boxed{SE}} \rightarrow N(0,1).$$

$\sqrt{i^{-1}(\hat{\beta})}$

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

- We can obtain one-sided ($\beta > \beta_0$) or two-sided ($\beta = \beta_0$) 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.

  $z^2 = \left(\frac{\hat{\beta} - \beta_0}{SE}\right)^2 \sim \chi^2(1)$. Under H_0
- This type of statistic is called the Wald statistic.

Wald test

Multivariate extension of the Wald test

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

The Wald statistic can be written as:

$$W = \underbrace{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})'}_{1 \times m}\underbrace{[\text{cov}(\hat{\boldsymbol{\beta}})]^{-1}}_{m \times m}\underbrace{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})}_{m \times 1.} \overset{\text{Under } H_0}{\sim} \chi^2(m)$$

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

Wald

The **Wald statistic**:

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

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

Usually

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

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

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

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 $\ell(\beta_1)$ is the maximized value over all parameter space. The ratio Λ cannot exceed 1.

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

$$-2\ln\Lambda \;=\; -2\ln\big(\ell(\beta_0)/\ell(\hat{\beta})\big) \;=\; -2\big(L(\beta_0) - L(\hat{\beta})\big) \;> 0.$$

$\sim \chi_1^2$

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

$$-2\ln\Lambda \;\geq\; \boxed{\chi_1^2(\alpha)}$$

Score

The **score statistic**:

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

(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**:

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

(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:

$$z_S^2 = \frac{u(\beta_0)^2}{\imath(\beta_0)} \underset{H_0}{\dot{\sim}} \chi_1^2$$

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

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