SS9055B: Generalized Linear Models

Section 3: Maximum Likelihood Inference Part 2

Objectives

This section continues the discussion of the methods of maximum likelihood inference. By the end of this section you should be able to:

1. derive Wald, likelihood ratio, and score test statistics for simple models, and
2. invert the test statistics to obtain approximate confidence intervals.

Once again, this material is theoretical and may be new to most of you. **IT'S STILL NOT TIME TO PANIC!** We will work through examples in class.

Introduction

In the previous section of notes we derived the approximate sampling distribution for maximum likelihood estimators. For a model with a one dimensional parameter, θ, and assuming the data are *iid*, we found that

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{\mathcal{D}} N\left(0, \mathcal{I}^{(1)}(\theta^*)^{-1}\right)$$

where $\hat{\theta}$ is the MLE, θ^* is the true parameter value, and $\mathcal{I}^{(1)}(\theta^*)$ is the information in one observation evaluated at the true parameter. In practice, this allows us to approximate the sampling distribution of $\hat{\theta}$ as

$$\hat{\theta} \dot{\sim} N\left(\theta^*, \frac{1}{n\mathcal{I}^{(1)}(\hat{\theta})}\right).$$

The equivalent expressions in p dimensions are

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathcal{D}} N_p\left(0, \mathcal{I}^{(1)}(\boldsymbol{\theta}^*)^{-1}\right)$$

and

$$\hat{\boldsymbol{\theta}} \dot{\sim} N_p\left(\boldsymbol{\theta}^*, \frac{1}{n}\mathcal{I}^{(1)}(\hat{\boldsymbol{\theta}})^{-1}\right).$$

Notice that we replace the true value of the parameter with the estimate in the expression for the asymptotic variance, but not the mean. In fact, if you wanted to approximate the sampling distribution of the MLE then you would want to replace both. However, we are going to need the true parameter to appear in the approximate distribution in order to construct test statistics and confidence intervals, which is what we will do now.

Review

The methods for obtaining tests statistics and inverting them to construct confidence sets may seem new at first, but I am certain that you have all seen versions of these tests and confidence intervals before. Recall that the least squares estimators of the coefficients in a linear regression model exactly follow a normal distribution if the residual variance, σ^2, is known. In particular

$$\hat{\beta}_j \sim N(\beta^*, \sigma_j^2).$$

where σ_j^2 is the $j-th$ entry along the diagonal of the variance covariance matrix, $\Sigma = \sigma^2(X'X)^{-1}$. The hypotheses

$$H_0 : \beta_j = 0 \text{ versus } H_a : \beta_j \neq 0$$

can then be compared with the z-statistic

$$z = \frac{\hat{\beta}_j}{\sigma_j}$$

which has a standard normal distribution if H_0 is true. The p-value for the test is

$$p = 2P(Z > |z|)$$

where $Z \sim N(0,1)$. If we set a significance level, α, then we reject H_0 if $|z| > z_{\alpha/2}$ where $P(Z > z_{\alpha/2}) = \alpha/2$.

To obtain a confidence interval for β_j, note that we fail to reject H_0 if $|\hat{\beta}_j| < \sigma_j z_{\alpha/2}$ or equivalently if $\hat{\beta}_j - \sigma_j z_{\alpha/2} < 0 < \hat{\beta}_j + \sigma_j z_{\alpha/2}$. This yields the $(1-\alpha)100\%$ confidence interval $(\hat{\beta}_j - \sigma_j z_{\alpha/2}, \hat{\beta}_j + \sigma_j z_{\alpha/2})$. The set of estimates for which we fail to reject (accept) H_0, $A = \{\hat{\beta} : |\hat{\beta} - 0| < \sigma_j z_{\alpha/2}\}$ is called the acceptance region of the test and the process of constructing the confidence interval from A is called inverting the test statistic.

Likelihood Based Hypothesis Tests

To make the discussion general I will consider that the data consist of a random sample from some distribution dependent on a vector of parameters, $\boldsymbol{\theta}$, of length p. I will then consider testing the hypothesis that some subset of $k < p$ of the parameters are equal to 0. Notationally, I will assume that the parameters are ordered so that we can divide the vector of parameters into two pieces, $\boldsymbol{\theta} = (\boldsymbol{\theta}_0', \boldsymbol{\theta}_1')'$, where $\boldsymbol{\theta}_0$ is of length k and contains the parameters of interest and $\boldsymbol{\theta}_1$ is of length $p - k$ and contains the remaining parameters. The hypotheses we wish to test are then

$$H_0 : \boldsymbol{\theta}_0 = \mathbf{0} \text{ versus } H_a : \boldsymbol{\theta}_0 \neq \mathbf{0}.$$

The null hypothesis indicates that all elements of $\boldsymbol{\theta}_0$ are equal to 0 and the alternative indicates that at least one of the coefficients in $\boldsymbol{\theta}_0$ is not equal 0.

It may seem restrictive to assume that the null hypothesis sets the parameters of interest equal to 0. However, there are two reasons for doing this. First, this is the test that we will be interested in most often in the class because we will usually be testing whether or not some predictor(s) are important in the model which is equivalent to testing whether their coefficients should be 0. Second, we can usually obtain a test of this form by reparameterizing the distribution. Suppose that we had a normal sample and wished to test that the hypothesis that $\mu = 42$. An equivalent test is to define $\mu = \theta + 42$ and then test whether $\theta = 0$.

Wald Test

The simplest tests for maximum likelihood inference are based on directly computing probabilities from the asymptotic normal distribution of the MLE. If $k = 1$ then the hypotheses become

$$H_0 : \theta = 0 \text{ versus } H_a : \theta \neq 0.$$

The Wald test in this case ends up being exactly equivalent to a standard z-test using the estimated mean and standard error of the parameter of interest and ignoring the correction for the estimation of σ^2 which leads to the t-test, exactly as in Section . Explicitly, the test statistic is

$$z = \hat{\theta}/\text{SE}(\hat{\theta})$$

and the p-value is

$$p = P(Z < -|z| \text{ or } Z > |z|) = 2P(Z > |z|)$$

where $Z \sim N(0, 1)$. Equivalently, since the square of a standard normal random variable is a chi-square random variable with one degree of freedom we can compute the p-value as

$$p = P(Z^2 > z^2) = P(X > z^2)$$

where $X \sim \chi^2_1$.

More generally, when $k > 1$ we can construct a test statistic that has an approximate chi-squared distribution with k degrees of freedom if the null hypothesis is true. Explicitly, the Wald test statistic for testing the hypotheses

$$H_0 : \boldsymbol{\theta}_0 = \mathbf{0} \text{ versus } H_a : \boldsymbol{\theta}_0 \neq \mathbf{0}$$

is

$$W = \hat{\boldsymbol{\theta}}_0' I^{(n)}_{[1:k,1:k]}\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}}_0$$

where $I^{(n)}_{[1:k,1:k]}(\hat{\boldsymbol{\theta}})$ represents the $k \times k$ submatrix of the information matrix associated with $\boldsymbol{\theta}_0$. Under the regularity conditions for maximum likelihood inference, W has an approximate chi-square distribution with k degrees of freedom, $W \xrightarrow{\mathcal{D}} \chi^2_k$. The p-value is then computed as $p \approx P(X > W)$ where $X \sim \chi^2_k$.

Note that once again we have replaced the unknown true parameter value with the point estimate, $\hat{\boldsymbol{\theta}}$, in the computation of the information matrix. Tests based on W and its approximate chi-square distribution are called Wald tests.

Likelihood Ratio Tests

The second class of tests based on the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is the likelihood ratio test (LRT). Heuristically, the LRT compares the likelihood of the data under both the null and alternative hypotheses. It is always possible to make the likelihood bigger under the alternative hypothesis because the likelihood will never be maximized exactly when $\boldsymbol{\theta}_0 = \mathbf{0}$ simply because of random variation in the sample. However, we might be convinced to believe that the null hypothesis is true if $\hat{\boldsymbol{\theta}}_0$ is close to $\mathbf{0}$. The LRT measures this difference by comparing the maximum of the likelihood under the null and alternative hypotheses is small.

First we consider the likelihood when the null hypothesis is true. If the null hypothesis is true then we have to set $\boldsymbol{\theta}_0 = \mathbf{0}$, and so the likelihood is maximized by varying $\boldsymbol{\theta}_1$ alone. Let $\hat{\boldsymbol{\theta}}_0$ denote the point at which the likelihood is maximized when $\boldsymbol{\theta}_0 = \mathbf{0}$ so that

$$\max_{\boldsymbol{\theta}_1} L((\mathbf{0}, \boldsymbol{\theta}_1)|\boldsymbol{y}) = L(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y}).$$

Next we consider the likelihood under the alternative hypothesis. The alternative hypothesis imposes no restrictions on the parameter values and so the likelihood is maximized by the full MLE,

$$\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y}) = L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}).$$

The likelihood ratio compares these two values and is defined as

$$\Lambda = \frac{\max_{\boldsymbol{\theta}_1} L((\mathbf{0}, \boldsymbol{\theta}_1)|\boldsymbol{y})}{\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y})} = \frac{L(\hat{\boldsymbol{\theta}}_0^{\dagger}|\boldsymbol{y})}{L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})}.$$

The distribution of this statistic is hard to work with directly; however, under the same regularity conditions that imply asymptotic normality of the MLE, which are satisfied by all of the models we will consider, the statistic

$$G^2 = -2 \log \Lambda = 2[\log L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \log L(\hat{\boldsymbol{\theta}}^\dagger|\boldsymbol{y})]$$

has an asymptotic chi-squared distribution with k degrees of freedom. This is called the likelihood ratio test statistics (even though its actually equal to -2 times the log of the likelihood ratio). The p-value for the likelihood ratio test can then be computed as $p \approx P(X > G^2)$ where $X \sim \chi^2_k$.

Comparison

Both the Wald and likelihood ratio tests can be applied to test the same hypotheses, and so it is natural to ask which should be used. In general, the LRT is more powerful than the Wald test meaning that it is more likely to reject the null hypothesis given that the alternative hypothesis is true. The LRT statistic also converges to the asymptotic chi-square distribution more quickly meaning that p-values from the LRT will be more accurate. However, the Wald test statistic is generally easier to compute, though we will see that the LRT has a simple general form for GLM. You will see both Wald tests and the LRT on a regular basis in this class and beyond.

Score Tests

There is one other common test based on maximum likelihood theory which is the score test. Score tests assess the distance between the MLE and the value of the parameter specified by the null hypothesis by comparing the gradient of the likelihood at these points. The advantage of these tests is that we only have to consider what happens under the null hypothesis because we know that the gradient is 0 at the maximum (provide the likelihood is regular). This can be very convenient if the likelihood is complicated or if k is large so that the full likelihood has much higher dimension than the likelihood under the null hypothesis. However, Wald and likelihood ratio tests are easily conducted for the types of models we will consider and so we will not encounter score tests.

Example

Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed exponential random variables with mean μ and we wish to test the hypotheses

$$H_0 : \mu = 1 \text{ versus } H_a : \mu \neq 1.$$

The density of the exponential distribution is

$$f(x;\mu) = \frac{1}{\mu}e^{-x/\mu}, \quad x > 0$$

and so the likelihood is

$$L(\mu;\boldsymbol{x}) = \frac{1}{\mu^n}e^{-n\bar{x}/\mu}.$$

Equivalently, we can define $\theta = \mu - 1$ so that $Y_1, \ldots, Y_n \sim Exp(\theta+1)$ and test the hypotheses

$$H_0 : \theta = 0 \text{ versus } H_a : \theta \neq 0.$$

The density of the exponential distribution in terms of the new parametrization is

$$f(x;\theta) = \frac{1}{\theta+1}e^{-x/(\theta+1)}, \quad x > 0$$

and so the likelihood is

$$L(\mu;\boldsymbol{x}) = \frac{1}{(\theta+1)^n}e^{-n\bar{x}/(\theta+1)}.$$

It's straightforward to show that the MLE of μ is $\hat{\mu} = \bar{X}$ and so $\hat{\theta} = \bar{X} - 1$ (a property called invariance). Moreover, applying the CLT we have that

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\mu^2}{n}\right)$$

so that

$$\hat{\theta} \dot{\sim} N\left(\mu - 1, \frac{\mu^2}{n}\right).$$

The Wald test statistic is

$$W = \left(\frac{\hat{\theta}}{SE(\hat{\theta})}\right)^2 = \left(\frac{\bar{x}-1}{\bar{x}/\sqrt{n}}\right)^2 = n\left(1 - \frac{1}{\bar{x}}\right)^2$$

and the p-value is $P(\chi^2_1 > W)$. the likelihood ratio is

$$\begin{aligned}
\Lambda &= \frac{L(\mu^\dagger;\boldsymbol{x})}{L(\hat{\mu};\boldsymbol{x})} \\
&= \frac{L(\mu^\dagger;\boldsymbol{x})}{L(\bar{x};\boldsymbol{x})} \\
&= \frac{(\mu^\dagger)^{-n}\exp\left(-\sum_{i=1}^n x_i/\mu^\dagger\right)}{\bar{x}^{-n}\exp\left(-n\sum_{i=1}^n x_i/\bar{x}\right)} \\
&= \left(\frac{\bar{x}}{\mu}\right)^n \exp\left[-n\left(\frac{\bar{x}}{\mu} - 1\right)\right].
\end{aligned}$$

The likelihood ratio test statistic is

$$G^2 = -2\log\Lambda = -2n\left[\log\left(\frac{\bar{x}}{\mu}\right) + \left(1 - \frac{\bar{x}}{\mu}\right)\right].$$

and the p-value is $P(\chi^2_1 > G^2)$.

Likelihood Based Confidence Intervals

Anytime we have a test statistic whose (approximate) distribution is known under the null hypotheis we can construct an (approximate) confidence set through the process of inversion. Given the null hypothesis $H_0 : \boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^\dagger$ and a level of significance, α, we can partition the sample space into two regions. The rejection region, $R(\boldsymbol{\theta}_0^\dagger)$, is the set of data points for which H_0 would be rejected at the α level of s0ignificance. The complement is the acceptance region, $A(\boldsymbol{\theta}_0^\dagger) = R(\boldsymbol{\theta}_0^\dagger)^C$, containing the data points for which H_0 would not be rejected (i.e., accepted). The process of inversion involves identifying the set of $\boldsymbol{\theta}_0^\dagger$ for which the acceptance region contains the observed test statistic.

The Wald and likelihood ratio tests both have rejection regions of the form:

$$R(\boldsymbol{\theta}_0^\dagger) = \{\boldsymbol{y} : T(\boldsymbol{y}, \boldsymbol{\theta}_0^\dagger) > \chi^2_{k,\alpha}\}$$

where $T(\boldsymbol{y}, \boldsymbol{\theta}_0^\dagger)$ is the test statistic computed given the observed data and the null hypothesis, k is the appropriate degrees of freedom, and $\chi^2_{k,\alpha}$ is the critical point such that $P(\chi^2_k > \chi^2_{k,\alpha}) = \alpha$. The corresponding acceptance region is:

$$A(\boldsymbol{\theta}_0^\dagger) = \{\boldsymbol{y} : T(\boldsymbol{y}, \boldsymbol{\theta}_0^\dagger) < \chi^2_{k,\alpha}\}.$$

Inverting this region gives the $(1-\alpha)100\%$ confidence set:

$$C(\boldsymbol{y}) = \{\boldsymbol{\theta}_0 : T(\boldsymbol{y}, \boldsymbol{\theta}_0) < \chi^2_{k,\alpha}\}.$$

Note that inversion is not guaranteed to produce a confidence set that comprises a single interval for all models However, this method will always produce an interval estimate for the three tests we have considered because the χ^2 distribution is unimodal.

Example

Consider the previous example. According to the CLT

$$\hat{\mu} = \bar{X} \dot{\sim} N\left(\mu, \frac{\mu^2}{n}\right).$$

The $(1-\alpha)100\%$ Wald confidence interval for μ is then given by

$$\bar{x} \pm z_{\alpha/2}\frac{\bar{X}}{\sqrt{n}}.$$

To find the $(1-\alpha)100\%$ confidence interval based on the LRT we first need to in invert the test statistic, G^2, found in the previous example. That is, we need to identify the values of $\mu^\dagger$ for which $G^2 < \chi^2_{1,\alpha}$. Unfortunately, this equation cannot be solved by hand and the endpoints must be computed numerically.

As an example, suppose that $n = 25$ and $\bar{x} = .95$. The endpoints of the 95% Wald confidence interval for μ would have endpoints

$$.95 - 1.96(.95/\sqrt{25}) = .578 \text{ and } .95 + 1.96(.95/\sqrt{25}) = 1.322.$$

The plot in Figure 1 show the value of the likelihood ratio test statistic as a function of μ. The horizontal line represents the value of $\chi^2_{1,.05} = 3.841$. Inverting the test statistic to construct the confidence interval is equivalent to finding all of the value of μ for which the test statistic is below this point. This is done by finding the two points where the lines cross, which are 0.6575306 and 1.444946. Hence, the 95% Wald confidence interval for μ is (.578,1.322) and the 95% likelihood ratio confidence interval is (.658,1.445).

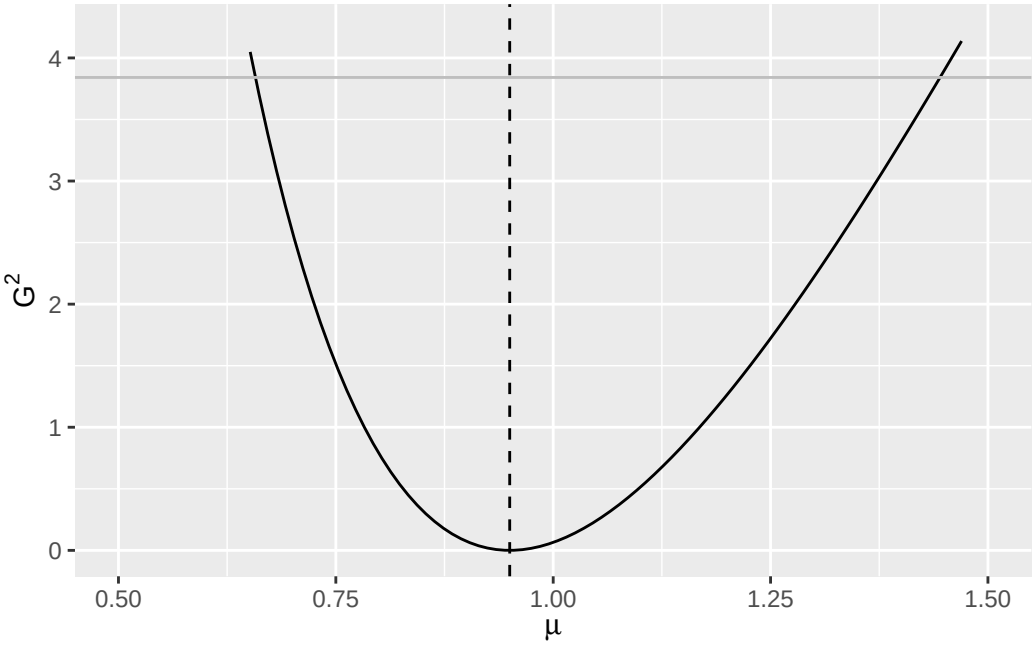

Figure 1: Likelihood ratio test statistic for a random sample of size 25 from an exponential distribution when $\bar{x} = .95$. The vertical dashed line represents the MLE (i.e., $\bar{x}$). The horizontal grey line represents the value $\chi^2_{1,.05} = 3.841$.