SS9055B: Generalized Linear Models

Winter 2024

Section 2: Maximum Likelihood Inference Part 1

Objectives

This section introduce the methods of maximum likelihood inference – the method of inference we will focus on in this course. By the end of this section you should be able to:

1. construct the likelihood function for a given model,
2. obtain the maximum likelihood estimators from simple likelihood functions,
3. construct the asymptotic distribution for these estimators, and
4. apply the delta method to make inference about functions of parameters.

Note that this material is theoretical and will be new to some of you. **DON'T PANIC!** We will work through examples in class, and you don't need to understand all of the theory to work with the applied material. However, I feel that it is important for you to have some theoretical understanding of the background behind GLM.

1 Introduction

In this course we will study some extensions of the linear regression framework. We will begin by looking at the class of generalized linear models (GLMs). The importance of this topic is that inference can be applied in a similar way to all models in the class. One need only show that a model fits the definition of a GLM and then all the tools of the framework become available. We will focus primarily on maximum likelihood inference and will begin with a general review of these methods. We will then describe the application of these tools to abstract GLMs and consider some particular cases including logistic regression for binomial data, Poisson log-linear regression for count data, and models for multinomial data. This section will also help to layout some of the basic notation for the course.

2 Likelihood, Point Estimates and the MLE

Least squares estimation is a very good procedure for the linear regression model. In fact, the least squares estimates are the best linear unbiased estimators. This means that if we find another unbiased estimator, $\tilde{\boldsymbol{\beta}}$, which is a linear combination of the response values, meaning that $\tilde{\boldsymbol{\beta}} = M\boldsymbol{Y}$ for some matrix M such that $E(M\boldsymbol{Y}) = \boldsymbol{\beta}$, then $\text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}})$ is a positive semidefinite matrix. This is the multidimensional equivalent of saying that $\text{Var}(\tilde{\beta}_j) - \text{Var}(\hat{\beta}_j) \geq 0$ in one dimension and is a result of the Gauss-Markov theorem.

Unfortunately, least squares estimation is not a very good procedure more generally. The more common approach in frequentist statistics, and the one we will consider throughout this entire course, is maximum likelihood estimation.

2.1 Likelihood Function

Suppose that our data consist of observations on n univariate random variables, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, having joint density (pmf or pdf) $f(\boldsymbol{y}|\boldsymbol{\theta})$ dependent on the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ of length p. I will denote the true value of the parameter by $\boldsymbol{\theta}^*$ and the observed value of $\boldsymbol{Y}$ by $\boldsymbol{y}$. The likelihood function is equal to the joint density of $\boldsymbol{Y}$ viewed as a function of the parameters given the observed data rather than as a function of the data given the parameters:

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

The log-likelihood function,

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \log L(\boldsymbol{\theta}|\boldsymbol{y}) = \log f(\boldsymbol{y}|\boldsymbol{\theta}),$$

provides the same information and is easier to work with for the distributions considered in this course which have an exponential form.

2.2 Maximum Likelihood Estimator

The standard point estimator for frequentist inference is the maximum likelihood estimator (MLE) which is commonly defined as the value of $\boldsymbol{\theta}$ which maximizes $L(\boldsymbol{\theta}|\boldsymbol{y})$:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}} L(\boldsymbol{\theta}|\boldsymbol{y})$$

where Θ represents the parameter space.

For most, if not all, of the models we consider the parameters will be continuous, the likelihood function will be log-concave, and the maximum will occur within the interior of the parameter space (i.e., not on the boundary). In this case, we can compute the MLE by solving the likelihood equations:

$$\frac{d}{d\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\frac{\partial l}{\partial \theta_1}, \ldots, \frac{\partial l}{\partial \theta_p}\right) = \mathbf{0}.$$

Note that you should always compute the Hessian, the matrix of second derivatives, to confirm that a maximum has been reached. The Hessian of the log-likelihood, $\boldsymbol{H}$, is the $p \times p$ matrix with j, k entry:

$$\frac{\partial^2}{\partial\theta_j \partial\theta_k} l(\boldsymbol{\theta}|\boldsymbol{y}).$$

The log-likelihood is concave and a maximum is reached if H is negative definite (i.e., if $\boldsymbol{x}'\boldsymbol{H}\boldsymbol{x} < 0$ for all $\boldsymbol{x}$). Checking this condition is not always easy.

In one dimension, the MLE is the value of θ which maximizes the likelihood function given the observed data:

$$\hat{\theta} = \underset{\theta\in\Theta}{\text{argmax}} L(\theta|\boldsymbol{y}).$$

This is usually found by solving the likelihood equation,

$$\frac{d}{d\theta} l(\theta|\boldsymbol{y}) = 0,$$

and then checking that the second derivative,

$$\frac{d^2}{d\theta^2} l(\theta|\boldsymbol{y})|_{\theta=\hat{\theta}},$$

is less than 0 to ensure that $\hat{\theta}$ is a maximum.

Example

Suppose that $Y_1, \dots, Y_n$ are independent normal random variables with unknown mean μ and variance 1. The density for a single observation is:

$$f(y_i|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu)^2}{2}\right)$$

and the joint density for all n observations is

$$f(\boldsymbol{y}|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i-\mu)^2}{2}\right) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2\right).$$

The likelihood function is formed simply by viewing $f(\boldsymbol{y}|\mu)$ as a function of μ not $\boldsymbol{y}$,

$$L(\mu|\boldsymbol{y}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2\right),$$

and the log-likelihood function is

$$l(\mu|\boldsymbol{y}) = \frac{-n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2.$$

Differentiating with respect to μ we find

$$\frac{dl}{d\mu} = \sum_{i=1}^n (y_i - \mu)$$

and

$$\frac{d^2 l}{d\mu^2} = -n.$$

Solving $dl/d\mu = 0$ yields $\mu = \sum_{i=1}^n y_i/n = \bar{y}$. Since $\frac{d^2 l}{d\mu^2} < 0$ for all values of μ this is the unique maximum. Hence, the MLE is $\hat{\mu} = \bar{y}$, the sample mean.

2.3 Maximum Likelihood Estimation for Linear Regression

As an example we will compute the maximum likelihood estimator for the linear regression model. The likelihood is

$$L(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right)$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution and the log-likelihood is

$$l(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = -n\log(\sigma) - \frac{\sum_{i=1}^n (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{\sigma^2} + C$$

where C is a constant that depends on neither $\boldsymbol{\beta}$ or σ^2. It's not important to know the exact value of C because the constant doesn't affect the location of the maximum or the derivatives with respect to the unknown parameters, which is all we will need. Note that for any fixed value of σ^2 the log-likelihood is maximized when the numerator of the second term,

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2,$$

is minimized. This occurs independent of σ^2 which means that the same value of $\boldsymbol{\beta}$ maximizes the likelihood for all values of σ^2 and must be the maximum likelihood estimator of $\boldsymbol{\beta}$. Moreover, the expression we are minimizing is simply the residual sum of squares, which shows that the MLE is exactly equal to the least squares estimator. This means that we can apply all of the methods we derived for least squares estimators above with one very minor change. The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

which is $(n-p-1)/n$ times the unbiased estimator. This estimator is biased, but has a smaller variance than the unbiased estimator. However, this usually makes little difference in practice since $(n-p-1)/n \approx 1$ unless the sample size is very small relative to the number of predictors, which would not be a good idea anyhow.

2.4 Fisher Information

An important quantity in maximum likelihood inference is the Fisher information matrix, denoted by $\mathcal{I}^{(n)}(\boldsymbol{\theta})$ and often referred to simply as the information. Given the setup in Section 2.1, the Fisher information matrix evaluated at $\boldsymbol{\theta}$ is the $p \times p$ matrix with i,j entry

$$\mathcal{I}_{i,j}^{(n)}(\boldsymbol{\theta}) = E_\theta\left[\left(\frac{\partial}{\partial\theta_i}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right)\right].$$

It is also possible to show that under certain conditions on the density, satisfied by all of the models we will consider, that

$$\mathcal{I}_{i,j}^{(n)}(\boldsymbol{\theta}) = -E_\theta\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right).$$

Heuristically, the information measures the average curvature of the likelihood function upon repeated sampling of the data. If the likelihood is very peaked on average then the second derivatives will be large (and negative) and the information will be big. In turn, this means that there is less uncertainty in the value of the parameters. If the curvature is low then the Fisher information will be small and the value of the parameters will be more uncertain.

Although this expression looks somewhat complicated there is an important simplification. If the elements of $\boldsymbol{Y}$ are *iid* then we can rewrite the density as a product of n terms:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}).$$

Note that I am being a little loose with notation by using $f(\cdot|\boldsymbol{\theta})$ to denote the density of both the vector $\boldsymbol{Y}$ and a single element $\boldsymbol{Y}_i$ and allowing the argument to indicate which variable is being considered. This is very common. Substituting this into the second expression for the information we then find that

$$\mathcal{I}_{i,j}^{(n)}(\boldsymbol{\theta}) = -E_\theta\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{i=1}^n \log f(y_i|\boldsymbol{\theta})\right) = -\sum_{i=1}^n E_\theta\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(y_i|\boldsymbol{\theta})\right) = n\mathcal{I}_{i,j}^{(1)}(\boldsymbol{\theta})$$

where $\mathcal{I}^{(1)}(\boldsymbol{\theta})$ represents the information from one observation (since the observations are identically distributed it does not matter which). This shows that the information in a sample of n *iid* observations is equal to n times the information in one observation.

Example

Consider the previous example in which $Y_1, \ldots, Y_n$ are iid normal random variables with unknown mean μ and fixed variance 1. We found that the second derivative of the log-likelihood with respect to μ was

$$\frac{d^2 l}{d\mu^2} = -n.$$

Hence, the Fisher information is

$$\mathcal{I}^{(n)}(\mu) = -E(-n) = n.$$

Note that $I^{(1)}(\mu) = 1$ so that $I^{(n)}(\mu) = nI^{(1)}(\mu)$.

2.4.1 Convergence in Distribution

The next section discusses how we approximate the sampling distribution of the MLE, which relies on the concept of convergence in distribution. You may not have seen this before if you have not studied advanced probability. Briefly, convergence in distribution means that as the sample size gets bigger the cumulative distribution function (CDF) of a random variable gets closer and closer to the CDF of the limit at (almost) every point in the sample space. In one dimension, suppose that $X_1, X_2, \ldots$ is a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots$. We say that the sequence converges in distribution to the random variable X if $F_n(x) \to F(x)$ wherever $F(x)$ is continuous. In practice, this means that $P(X_n \leq x) = F_n(x) \approx F(x) = P(X \leq x)$ so that we can approximate probabilities of events defined in terms of X_n with probabilities for the same events written in terms of X when n is large. This is written in shorthand as $X_n \overset{\mathcal{D}}{\to} X$. The definition in multiple dimensions is similar so that the sequence of random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ converges in distribution to the random vector X if the joint CDFs converge pointwise. Again, we write $\boldsymbol{X}_n \overset{\mathcal{D}}{\to} \boldsymbol{X}$ and approximate probabilities about $\boldsymbol{X}_n$ with probabilities about $\boldsymbol{X}$ when n is large enough.

Whether or not you know it you have all seen an example of convergence in distribution – the central limit theorem. The version of the central limit theorem that is taught in introductory statistics classes states that if $X_1, \ldots, X_n$ are iid random variables with mean μ and variance $\sigma^2 < \infty$ then the sample mean, $\bar{X}_n$ is approximately normal regardless of the original distribution provided that n is large enough. Mathematically, we write that

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

and usually say that n should be greater than 30 for the approximation to be accurate. But what does "approximately normal" mean. Approximately normal means that the distribution of $\bar{X}_n$ converges to a normal distribution. However, we have to be a little careful about what this really mean. In fact, $\bar{X}_n$ does not converge to a normal distribution because its variance, σ^2/n, converges to 0 as $n \to \infty$. If we were simply to look at the distribution of $\bar{X}_n$ as $n \to \infty$ we'd find that it converges to a point mass right at μ. However, if we used this to approximate the distribution of $\bar{X}$ then we would say that $\text{SE}(\bar{X}) = 0$, which is not very useful.

In order to approximate a distribution of $\bar{X}_n$ we need to multiply by a constant that stabilizes the variance (i.e., find c such that $c\bar{X}_n$ converges to some constant). The correct constant is $c = \sqrt{n}$ since $\text{Var}(\bar{X}_n) = \sigma^2/n$ and $\text{Var}(c\bar{X}_n) = c^2\text{Var}(\bar{X}_n)$. However, there is one final wrinkle. Simply multiplying $\bar{X}_n$ by $\sqrt{n}$ stabilizes the variance, but the mean of $\sqrt{n}\bar{X}_n$ converges to ∞. To avoid this we first have to subtract the mean, μ. This is what leads us to consider the distribution of $\sqrt{n}(\bar{X}_n - \mu)$ and to the fact that

$$\sqrt{n}(\bar{X}_n - \mu) \overset{\mathcal{D}}{\to} N(0, \sigma^2).$$

Note that you will be able to understand the material in this course without fully understanding the concept of convergence in distribution, so you do not need to get hung up on this. It is sufficient to know that $\bar{X}_n \dot{\sim} N(\mu, \sigma^2/n)$, but I think it is helpful to have some idea of what this really means.

2.5 Approximate Sampling Distribution

The reason that the Fisher information is important is that it provides an approximation to the variance of the MLE that we will need to computed standard errors, construct confidence intervals, and conduct hypothesis tests. More specifically, the Fisher information is needed to approximate the variance of the sampling distribution of the MLE. We know that the MLE for $\boldsymbol{\beta}$ in the linear regression model must have a normal sampling distribution since it is equal to the least squares estimator which is a linear transformation of the data. This is not true in general, the sampling distribution of the MLE will rarely be exactly normal, but it is possible to show that the MLE is approximately normal under fairly general conditions provided that the sample size is large. By this we mean that the MLE, appropriately scaled so that the variance does not shrink to 0, converges in distribution to a multivariate normal distribution.

Suppose that the elements of $\boldsymbol{Y}$ form an *iid* sample of size n from some univariate density so that the likelihood function is:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}).$$

Under regularity conditions, which we will again assume are satisfied, the ML estimator is an asymptotically normal estimator such that

$$\mathcal{I}^{(n)}(\boldsymbol{\theta}^*)^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \overset{\mathcal{D}}{\to} N_p\left(0, I_p\right)$$

where $\boldsymbol{\theta}^*$ is the true value of the parameter and I_p is the identity matrix of dimension p. Since the observations are assumed to be *iid* we can set $\mathcal{I}^{(n)}(\boldsymbol{\theta}^*) = n\mathcal{I}^{(1)}(\boldsymbol{\theta}^*)$ to obtain

$$\sqrt{n}\mathcal{I}^{(1)}(\boldsymbol{\theta}^*)^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \overset{\mathcal{D}}{\to} N_p\left(0, I_p\right).$$

If we multiply the left by $\mathcal{I}^{(1)}(\boldsymbol{\theta}^*)^{-1/2}$ then we find that

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

The practical application of this is that if n is large enough then we approximate the sampling distribution of $\hat{\boldsymbol{\theta}}$ with a multivariate normal distribution:

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

In particular, this implies that the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be approximated by the inverse information matrix,

$$\text{Var}(\hat{\boldsymbol{\theta}}) \approx \frac{1}{n}\mathcal{I}^{(1)}(\boldsymbol{\theta}^*)^{-1}.$$

The approximate standard deviation of $\hat{\theta}_j$ is given by the square root of the j,j element of this matrix. Since we don't know the true parameter values we replace $\boldsymbol{\theta}^*$ with $\hat{\boldsymbol{\theta}}$ to obtain the further approximation:

$$\widehat{\text{Var}(\hat{\boldsymbol{\theta}})} \approx \frac{1}{n}\mathcal{I}^{(1)}(\hat{\boldsymbol{\theta}})^{-1}.$$

The square root of the j^{th} diagonal element of this matrix is called the standard error of $\hat{\theta}_j$. Note that the standard error involves two separate levels of approximation. The first occurs when we approximate the true distribution of the estimator by the normal distribution. The second occurs when we replace the unknown parameters by their estimates.

More generally, we can relax the assumption that the random variables are identically distributed provided that other conditions are satisfied. The exact result in this case is more difficult to state because the information for a single observation is not constant. However, the final result is similar. Under regularity conditions and given that n is large enough, we can approximate the sampling distribution of $\hat{\boldsymbol{\theta}}$ by:

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

We can then approximate the variance matrix with

$$\widehat{\text{Var}(\hat{\boldsymbol{\theta}})} \approx \mathcal{I}^{(n)}(\hat{\boldsymbol{\theta}})^{-1}.$$

and the square root of the j^{th} diagonal element of this matrix is again called the standard error of $\hat{\theta}_j$.

Example

Considering the earlier example, we found that $I^{(n)}(\mu) = n$. Hence, $\text{Var}(\hat{\mu}) = 1/n$ and the approximate standard error of $\hat{\mu}$ is $\text{SE}(\hat{\mu}) = 1/\sqrt{n}$.

Note that this should not have been a surprise. Recall that the MLE is the sample mean, $\hat{\mu} = \bar{y}$. In this case $\hat{\mu}$ is exactly normal and the standard deviation of $\hat{\mu}$ is exactly $1/\sqrt{n}$.

3 Further Considerations

3.1 Performance of the MLE

The performance of different estimators for the same parameters are generally compared in terms of their bias, standard error, and mean-squared error (equal to the square of the bias plus the square of the standard error). The MLE is not necessarily unbiased and does not have the smallest variance of all possible estimators, in general. However, use of the MLE is supported by asymptotic arguments – i.e., by its performance when the sample size is large. Under the same regularity conditions that provide asymptotic normality of MLEs, the MLE is both consistent and asymptotically efficient. Consistency means that the MLE converges in probability to the true parameter value, denoted by $\hat{\boldsymbol{\theta}} \overset{\mathcal{P}}{\to} \boldsymbol{\theta}^*$. Practically speaking, this means the estimator is likely to be very close the true parameter value when n is large. Asymptotic efficiency means that the variance of the asymptotic distribution of the MLE is as small as possible for any asymptotically normal estimator. This essentially means that the MLE will have the smallest variance of all aysmptotically unbiased estimators. Combined, these mean that the MLE is guaranteed to be the best estimator when the sample size is very, very large – and are likely to be good estimators when the sample size is reasonably large. This is why maximum likelihood estimation is applied in so many problems.

3.2 Delta Method

Finally, we will look at one more result called the delta method. Sometimes we will interested in making inference about some function of the parameters in a model instead of the parameters themselves. To do this we need to understand the sampling distribution of a function of the MLE. The delta method provides a way to do because the MLE has an approximate normal distribution. Although statisticians give it a special name, the delta method is nothing more than an application of first order Taylor series expansion.

Strictly speaking, suppose that $\boldsymbol{X}_n$ is a random vector of length p which has an asymptotically normal distribution

$$\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\mu}) \overset{\mathcal{D}}{\to} N_p(0, \Sigma)$$

for some mean vector $\boldsymbol{\mu}$ of length p and some variance matrix, Σ, of dimension $p \times p$. Let $g(\boldsymbol{x})$ be a differentiable, vector valued function mapping $\Re^n$ to $\Re^m$ and define $\boldsymbol{Y}_n = g(\boldsymbol{X}_n)$. Then:

$$\sqrt{n}(\boldsymbol{Y}_n - g(\boldsymbol{\mu})) \overset{\mathcal{D}}{\to} N\left(0, J(\boldsymbol{\mu})\Sigma J(\boldsymbol{\mu})'\right)$$

where $J(\boldsymbol{\mu})$ is the Jacobian – the matrix of partial derivatives of $g(\boldsymbol{x})$ evaluated at $\boldsymbol{\mu}$ such that

$$J_{i,j} = \left.\frac{\partial g_i}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{\mu}}$$

where $g_i(\boldsymbol{x})$ represents the i^{th} element of the output of $g(\boldsymbol{x})$. Applying this to the MLE for $\boldsymbol{\theta}$ and assuming that the observations are *iid* implies that

$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}^*)) \overset{\mathcal{D}}{\to} N(0, J(\boldsymbol{\theta}^*)\mathcal{I}^{-1}(\boldsymbol{\theta}^*)J(\boldsymbol{\theta}^*)').$$

Since the true value of the parameters, $\boldsymbol{\theta}^*$, is unknown, we must replace it with the MLE in practice.

Practically speaking, this means that we can also approximate the sampling distribution of $g(\hat{\boldsymbol{\theta}})$ by a normal distribution. Specifically,

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

We can then use this result to obtain approximate hypothesis tests and confidence regions for $\boldsymbol{\theta}$ without having to recompute the distribution of $g(\boldsymbol{\theta})$ directly. If $\boldsymbol{\theta}$ is univariate, $p = 1$, then we get the simpler result

$$\sqrt{n}(g(\hat{\theta}) - g(\theta^*)) \xrightarrow{\mathcal{D}} N\left(0, \frac{[g'(\theta^*)]^2}{\mathcal{I}^{(1)}(\theta^*)}\right)$$

which gives the approximation

$$g(\hat{\theta}) \dot{\sim} N\left(g(\theta^*), \frac{[g'(\hat{\theta})]^2}{n\mathcal{I}^{(1)}(\hat{\theta})}\right)$$

when θ^* is replaced by the MLE. This implies that the standard error of $g(\hat{\theta})$ is approximately equal to $|g'(\hat{\theta})| \Big/ \sqrt{n\mathcal{I}^{(1)}(\hat{\theta})}$.

Example

Consider the previous example again and suppose that we were actually interested in estimating the value e^{μ}. If we let $g(x) = e^x$ then $g'(x) = e^x$. Applying the delta method we get that

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

Hence, the standard error of $e^{\bar{y}}$ is $SE(e^{\bar{y}}) = e^{\bar{y}}/\sqrt{n}$.

4 Summary

In this course we fill focus primarily on maximum likelihood methods to estimate parameters and conduct inference by computing confidence intervals and completing hypothesis tests. In the special case of the linear regression model maximum likelihood inference coincides with least squares inference. As a result, the maximum likelihood estimator has an exact normal distribution form which confidence intervals are easily computed and the sampling distribution of test statistics are simple to derive. While this is not true more generally, the models we will consider all satisfy the regularity conditions so that the sampling distribution of the estimator is approximately normal and the same methods can be applied to obtain approximate confidence intervals and perform approximate hypothesis tests. In the next section we will consider some further methods of inference based on the likelihood before we consider the framework of generalized linear models.