Maximum Likelihood 1

Jacob M. Montgomery

2017

Maximum likelihood estimation

Overview

- In this class we will talk about point estimates from four perspectives
  - Frequentist
  - Maximum likelihood
  - Bayesian
  - Nonparametric
- Today we are going to talk about maximum likelihood estimation
  - What is an MLE estimate?
  - What are the properties of these estimators?

Big picture

We are trying to estimate a parameteric model

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$$

where $\Theta \subset R^k$.

Big picture

We are trying to estimate a parameteric model

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$$

where $\Theta \subset R^k$.

- We have made assumptions about the DGP that allows us to write out a formula.

Big picture

We are trying to estimate a parameteric model

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$$

where $\Theta \subset R^k$.

- We have made assumptions about the DGP that allows us to write out a formula.
- If we can just estimate $\theta = (\theta_1, \ldots, \theta_n)$ we can fully charaterize the DGP.

Interest and nuissance

- **Parameters of interest** are the parameters we actually care about to make our point.

Interest and nuisance

- **Parameters of interest** are the parameters we actually care about to make our point.
- **Nuisance parameters** are parameters that we don't care about, but which still are necessary to characterize the dgp.

Interest and nuissance

- **Parameters of interest** are the parameters we actually care about to make our point.
- **Nuisance parameters** are parameters that we don't care about, but which still are necessary to characterize the dgp.

Example:

$$Y \sim N(\mathbf{X}\beta, \sigma^2)$$

- We need to estimate $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$.
- We usually don't care much about σ^2.

Method of moments

1. We are going to define the j^{th} moment as:

$$\alpha_j \equiv \alpha_j(\theta) = E(X^j) = \int x^j f(x) dx$$

Method of moments

1. We are going to define the j^{th} moment as:

$$\alpha_j \equiv \alpha_j(\theta) = E(X^j) = \int x^j f(x) dx$$

2. We assume that the j^{th} moment is equivalent to the j^{th} **sample moment** defined as:

$$\hat{\alpha}_j \equiv \hat{\alpha}_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

Method of moments

1. We are going to define the j^{th} moment as:

$$\alpha_j \equiv \alpha_j(\theta) = E(X^j) = \int x^j f(x) dx$$

2. We assume that the j^{th} moment is equivalent to the j^{th} **sample moment** defined as:

$$\hat{\alpha}_j \equiv \hat{\alpha}_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

3. Recall that we can use the first j moments to calculate the first j **central moments**.

Method of moments

1. We are going to define the j^{th} moment as:

$$\alpha_j \equiv \alpha_j(\theta) = E(X^j) = \int x^j f(x)dx$$

2. We assume that the j^{th} moment is equivalent to the j^{th} **sample moment** defined as:

$$\hat{\alpha}_j \equiv \hat{\alpha}_j(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

3. Recall that we can use the first j moments to calculate the first j **central moments**.
4. Set the equations as a system of j equations with j unknowns and solve

Example:

Let $X_1, \ldots, X_n \sim$ *Normal*(μ, σ^2). *Use the methods of moments to estimate the model.*

Example:

Let $X_1, \ldots, X_n \sim$ *Normal*(μ, σ^2)*. Use the methods of moments to estimate the model.*

- $\alpha_1 = E(X^1) = \mu$

Example:

Let $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$. Use the methods of moments to estimate the model.

- $\alpha_1 = E(X^1) = \mu$
- $\alpha_2 = E(X^2)$

Example:

Let $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$. Use the methods of moments to estimate the model.

- $\alpha_1 = E(X^1) = \mu$
- $\alpha_2 = E(X^2)$
- $\sigma^2 = E[(X - \mu)^2] = E(X^2) - [E(X)]^2$

Example:

Let $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$. Use the methods of moments to estimate the model.

- $\alpha_1 = E(X^1) = \mu$
- $\alpha_2 = E(X^2)$
- $\sigma^2 = E[(X-\mu)^2] = E(X^2) - [E(X)]^2$
- Thus, $\hat{\mu} = \hat{\alpha}_1 = \bar{X}$,

Example:

Let $X_1, \ldots, X_n \sim$ Normal(μ, σ^2). Use the methods of moments to estimate the model.

- $\alpha_1 = E(X^1) = \mu$
- $\alpha_2 = E(X^2)$
- $\sigma^2 = E[(X - \mu)^2] = E(X^2) - [E(X)]^2$
- Thus, $\hat{\mu} = \hat{\alpha}_1 = \bar{X}$,
- and

$$\hat{\sigma}^2 = \hat{\alpha}_2 - \hat{\alpha}_1^2$$

Example:

Let $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$. Use the methods of moments to estimate the model.

- $\alpha_1 = E(X^1) = \mu$
- $\alpha_2 = E(X^2)$
- $\sigma^2 = E[(X-\mu)^2] = E(X^2) - [E(X)]^2$
- Thus, $\hat{\mu} = \hat{\alpha}_1 = \bar{X}$,
- and

$$\hat{\sigma}^2 = \hat{\alpha}_2 - \hat{\alpha}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Properties of MM estimators

1. The estimate $\hat{\theta}$ usually exists (although recall that many distributions don't even have moments).

Properties of MM estimators

1. The estimate $\hat{\theta}$ usually exists (although recall that many distributions don't even have moments).
2. The estimate is consistent
3. The estimate is asymptotically Normal, meaning that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $N(0, \Sigma)$ where

$$\Sigma = g(E(YY'))g'$$

, $Y = (X, X^2, \ldots, X^k)', g = (g_1, \ldots, g_k)$ and $g_j = \frac{\partial a_j^{-1}(\theta)}{\partial \theta}$.

Properties of MM estimators

1. The estimate $\hat{\theta}$ usually exists (although recall that many distributions don't even have moments).
2. The estimate is consistent
3. The estimate is asymptotically Normal, meaning that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $N(0, \Sigma)$ where

$$\Sigma = g(E(YY'))g'$$

   , $Y = (X, X^2, \ldots, X^k)', g = (g_1, \ldots, g_k)$ and $g_j = \frac{\partial a_j^{-1}(\theta)}{\partial \theta}$.
4. Since this can be a pain in the ass to calculate, we will often use other methods (e.g., parametric bootstraps) to estimate the uncertainty.

Maximum likelihood estimation

Maximum likelihood estimation (Fisher 1922, 1925) is a classic method that finds the value of the estimator "most likely to have generated the observed data, **assuming the model specification is correct**."

MLE as a logical procedure: Closed form solution

1. Calculate the likelihood for your data.

MLE as a logical procedure: Closed form solution

1. Calculate the likelihood for your data.
2. Take the log.

MLE as a logical procedure: Closed form solution

1. Calculate the likelihood for your data.
2. Take the log. Why?

MLE as a logical procedure: Closed form solution

1. Calculate the likelihood for your data.
2. Take the log. Why?
3. Take the first derivative, set equal to zero, and solve.

MLE as a logical procedure: Closed form solution

1. Calculate the likelihood for your data.
2. Take the log. Why?
3. Take the first derivative, set equal to zero, and solve.
4. Check your second order conditions if necessary.

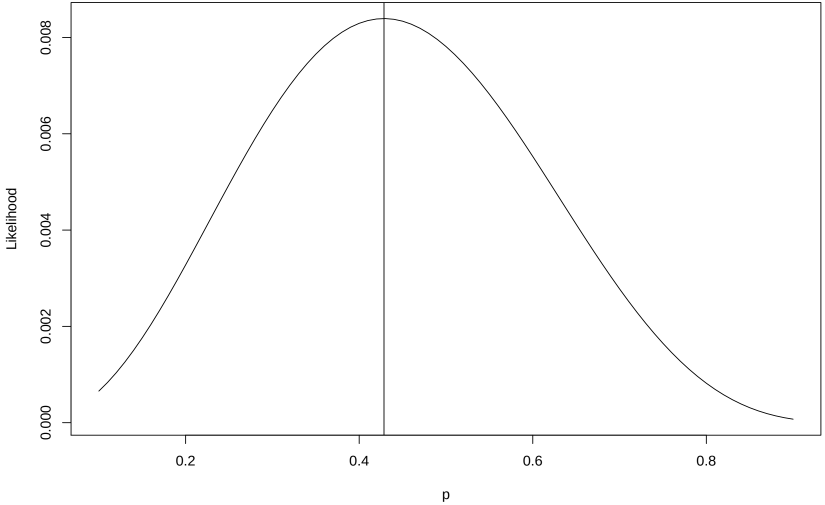
MLE as a logical procedure: Closed form solution

1. Calculate the likelihood for your data.
2. Take the log. Why?
3. Take the first derivative, set equal to zero, and solve.
4. Check your second order conditions if necessary.
5. Profit

Example 1:

Suppose that $X_1, \ldots, X_n \sim Bern(p)$. *Find the MLE for* p.

```r
x<-c(0, 0, 1, 1, 0, 0, 1); S<-sum(x); n<-length(x)
L.theta<-function(p){p^S*(1-p)^(n-S)}
plot(seq(.1, .9, by=.01), L.theta(seq(.1, .9, by=.01)), type="l",
     xlab="p", ylab="Likelihood")
abline(v=S/n)
```



MLE Comments

- In many cases, closed form solutions will not exist.

MLE Comments

- In many cases, closed form solutions will not exist.
- So we will rely on numerical methods.

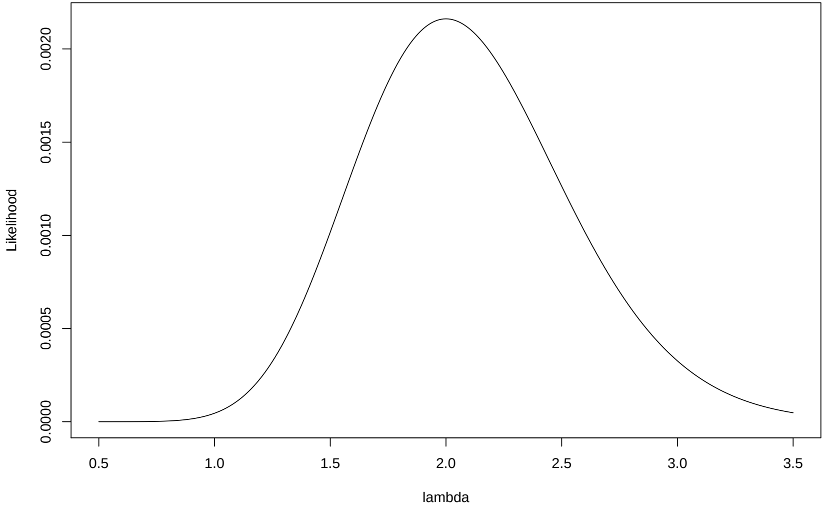
MLE Comments

- In many cases, closed form solutions will not exist.
- So we will rely on numerical methods.
- Since we are taking the derivative, any constant term (not involving θ) can be safely dropped.

Example 2:

Let $\mathbf{x} = \{5, 1, 1, 1, 0, 0, 3, 2, 3, 4\}$ *be data generated from a Pois*(λ) *distribution. Find the MLE for* λ.

```
x<-c(5,1, 1, 1, 0, 0, 3,2,3,4); S<-sum(x); n<-length(x)
L.theta<-function(lambda){exp(-n*lambda)*lambda^S}
plot(seq(.5, 3.5, by=.01), L.theta(seq(.5, 3.5, by=.01)), type="l",
     xlab="lambda", ylab="Likelihood")
```



When in doubt . . . can estimate with R . . . . but carefully

```
x<-c(5,1, 1, 1, 0, 0, 3,2,3,4); S<-sum(x); n<-length(x)
L.theta<-function(lambda){exp(-1*n*lambda)*lambda^S}
optim(par=4, fn=L.theta, method="BFGS")
```

```
## $par
## [1] 5.445418
##
## $value
## [1] 1.178897e-09
##
## $counts
## function gradient
##       17       16
##
## $convergence
## [1] 0
##
## $message
## NULL
```

When in doubt . . . can estimate with R . . . . but carefully

```
x<-c(5,1, 1, 1, 0, 0, 3,2,3,4); S<-sum(x); n<-length(x)
L.theta<-function(lambda){exp(-1*n*lambda)*lambda^S}
optim(par=.5, fn=L.theta, method="BFGS")
```

```
## $par
## [1] 0.3959845
##
## $value
## [1] 1.713268e-10
##
## $counts
## function gradient
##        8        7
##
## $convergence
## [1] 0
##
## $message
## NULL
```

When in doubt ... can estimate with R .... but carefully

```
x<-c(5,1, 1, 1, 0, 0, 3,2,3,4); S<-sum(x); n<-length(x)
L.theta<-function(lambda){exp(-1*n*lambda)*lambda^S}
optim(par=2, fn=L.theta, method="BFGS")
```

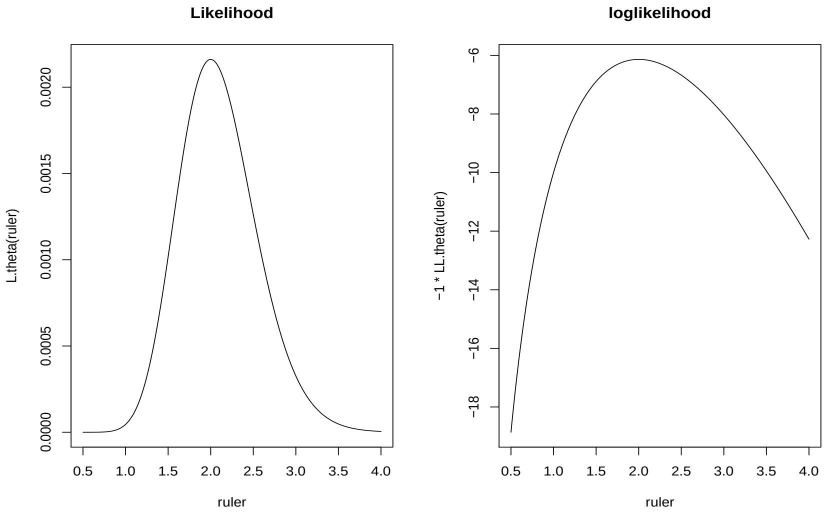
```
## $par
## [1] 2
##
## $value
## [1] 0.002161276
##
## $counts
## function gradient
##        4        1
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Use the log-likelihood

```r
x<-c(5,1, 1, 1, 0, 0, 3,2,3,4); S<-sum(x); n<-length(x)
LL.theta<-function(lambda){-1*(S*log(lambda) - n*lambda)}
optim(par=.5, fn=LL.theta, method="BFGS")
```

```
## $par
## [1] 2
##
## $value
## [1] 6.137056
##
## $counts
## function gradient
##       15        6
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
ruler<-seq(.5, 4, by=.01)
par(mfrow=c(1,2))
plot(x=ruler, y=L.theta(ruler), main="Likelihood", type="l")
plot(ruler, -1*LL.theta(ruler), main="loglikelihood", type="l")
```



Example 3:

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Find the MLE for $\theta = (\mu, \sigma^2)$

1. Start with μ
2. Go to next slide for σ.

We are going to substitute in $\mu = \bar{x}$ and drop irrelevant constants.

$$\frac{\partial}{d(\sigma^2)}\mathcal{L}(\mu, \sigma|\mathbf{x}) \propto \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

We are going to substitute in $\mu = \bar{x}$ and drop irrelevant constants.

$$\frac{\partial}{d(\sigma^2)}\mathcal{L}(\mu, \sigma|\mathbf{x}) \propto \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

We are going to substitute in $\mu = \bar{x}$ and drop irrelevant constants.

$$\frac{\partial}{d(\sigma^2)}\mathcal{L}(\mu, \sigma|\mathbf{x}) \propto \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$\frac{-n}{2\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

We are going to substitute in $\mu = \bar{x}$ and drop irrelevant constants.

$$\frac{\partial}{d(\sigma^2)}\mathcal{L}(\mu, \sigma|\mathbf{x}) \propto \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$\frac{-n}{2\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

$$-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

We are going to substitute in $\mu = \bar{x}$ and drop irrelevant constants.

$$\frac{\partial}{d(\sigma^2)}\mathcal{L}(\mu,\sigma|\mathbf{x}) \propto \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

$$= \frac{\partial}{d(\sigma^2)}\left(-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

$$\frac{-n}{2\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$$

$$-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Properties of MLE

1. We have already observed that the MLE is not necessarily unbiased

Properties of MLE

1. We have already observed that the MLE is not necessarily unbiased
2. The MLE is consistent

Properties of MLE

1. We have already observed that the MLE is not necessarily unbiased
2. The MLE is consistent
3. The MLE is asymptotically normal with standard errors easy(ish) to compute

Properties of MLE

1. We have already observed that the MLE is not necessarily unbiased
2. The MLE is consistent
3. The MLE is asymptotically normal with standard errors easy(ish) to compute
4. The MLE is asymptotically efficient.

Properties of MLE

1. We have already observed that the MLE is not necessarily unbiased
2. The MLE is consistent
3. The MLE is asymptotically normal with standard errors easy(ish) to compute
4. The MLE is asymptotically efficient.

These four facts together tell us the *exact* form of the asymptotic distribution for θ. What is it?

Uncertainty of the MLE

- The first derivative measures slope and the second derivative measures curvature of the function at a given point.

Uncertainty of the MLE

- The first derivative measures slope and the second derivative measures curvature of the function at a given point. - We already established that we use the first derivative to find the maximum.
  - The second derivative gives you information about the **rate of change** (acceleration).
  - The more peaked the function (the greater rate of change) at the MLE, the more “certain” the data are about the estimator.

Poisson example

We established above that:

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i$$

Poisson example

We established above that:

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i$$

From here we can calculate:

$$\frac{\partial^2 \mathcal{L}(\lambda)}{\partial \lambda^2} =$$

Poisson example

We established above that:

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i$$

From here we can calculate:

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

Asymptotic distribution

We know that as $n \to \infty$, $\hat{\theta}$ converges in distribution to:

$$\hat{\theta} \sim N\left(\theta, \frac{1}{I(\theta)}\right)$$

$$(\hat{\theta} - \theta) \sim N\left(0, \frac{1}{I(\theta)}\right)$$

Asymptotic distribution

We know that as $n \to \infty$, $\hat{\theta}$ converges in distribution to:

$$\hat{\theta} \sim N\left(\theta, \frac{1}{I(\theta)}\right)$$

$$(\hat{\theta} - \theta) \sim N\left(0, \frac{1}{I(\theta)}\right)$$

This second term is usally referred to as the **standard error** and is the standard deviation of the asymptotic distribution of $\hat{\theta}$.

- Of course we don't know either θ or $I(\theta)$ precisely.

- Of course we don't know either θ or $I(\theta)$ precisely.
- So, since we are in asymptopia anyways, we will simply **replace** $I(\theta)$ with $I(\hat{\theta})$.

- Of course we don't know either θ or $I(\theta)$ precisely.
- So, since we are in asymptopia anyways, we will simply **replace** $I(\theta)$ with $I(\hat{\theta})$.
- If that seems like a strong assumption, you are not alone.

Back to the poisson example

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

Now we need to calculate:

$$-E\left(-\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i\right)$$

Back to the poisson example

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

Now we need to calculate:

$$-E\left(-\frac{1}{\lambda^2} \sum_{i=1}^{n} x_i\right)$$

$$\frac{1}{\lambda^2} \sum_{i=1}^{n} \int x f(x) d_x$$

Back to the poisson example

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

Now we need to calculate:

$$-E\left(-\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i\right)$$

$$\frac{1}{\lambda^2}\sum_{i=1}^{n}\int xf(x)d_x$$

$$\frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

But we don't know λ, so what do we do?

But we don't know λ, so what do we do?

$$Var(\hat{\lambda}) = \frac{1}{I(\lambda)} = \frac{\lambda}{n} = \frac{\hat{\lambda}}{n}$$

But we don't know λ, so what do we do?

$$Var(\hat{\lambda}) = \frac{1}{I(\lambda)} = \frac{\lambda}{n} = \frac{\hat{\lambda}}{n} = \frac{\bar{x}}{n}$$

Jargon Alert!!!

- The **score function** is the first derivative of the log-likelihood:

$$S(\theta) = \frac{d}{d\theta}\mathcal{L}(\theta)$$

Jargon Alert!!!

- The **score function** is the first derivative of the log-likelihood:

$$S(\theta) = \frac{d}{d\theta}\mathcal{L}(\theta)$$

- The negative of the second derivative (giving curvature) of the log-likelihood is called the **observed information** or (sometimes to confuse you) the **observed Fisher's information**.

$$-\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta)$$

- The expected value of the second derivative of the log-likelihood is called the **Fisher information**, the **expected information**, or **expected Fisher's information**.

$$E\left[-\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta)\right]$$

MLE for multivariate models

- So it turns out we are often interested in estimating more than one parameter at the same time.

MLE for multivariate models

- So it turns out we are often interested in estimating more than one parameter at the same time.
- Just like we can extend the concepts of optimization (minimization) from single-variable calculus to multivariate calculus, we can find the MLE for a vector valued parameter $\theta = (\theta_1, \ldots, \theta_k)$.

MLE for multivariate models

- So it turns out we are often interested in estimating more than one parameter at the same time.
- Just like we can extend the concepts of optimization (minimization) from single-variable calculus to multivariate calculus, we can find the MLE for a vector valued parameter $\theta = (\theta_1, \ldots, \theta_k)$.
- **Parameters**: $\theta = (\theta_1, \ldots, \theta_k)$
- **Estimates**: $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$

MLE for multivariate models

- So it turns out we are often interested in estimating more than one parameter at the same time.
- Just like we can extend the concepts of optimization (minimization) from single-variable calculus to multivariate calculus, we can find the MLE for a vector valued parameter $\theta = (\theta_1, \ldots, \theta_k)$.
- **Parameters**: $\theta = (\theta_1, \ldots, \theta_k)$
- **Estimates**: $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$
- $\mathcal{L}(\theta) = \mathcal{L} = \sum_i^n \ln f(X_i|\theta)$

- Recall that the equivalent of a first derivative is termed the *gradient*.

$$\nabla \mathcal{L}(\theta)$$

- The multi-demensional extension of the second derivative is the *Hessian*.

$$H\left(\mathcal{L}(\theta)\right)$$

- Let $H_{jj} = \frac{\partial^2 \mathcal{L}}{\partial \theta_j^2}$
- Let $H_{jk} = \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k}$

- Let $H_{jj} = \frac{\partial^2 \mathcal{L}}{\partial \theta_j^2}$
- Let $H_{jk} = \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k}$
- The **Expected Fisher Information Matrix** is then the expected value of the Hessian.

$$I(\theta) = \begin{bmatrix} E(H_{11}) & E(H_{12}) & \dots & E(H_{1k}) \\ E(H_{21}) & E(H_{22}) & \dots & E(H_{2k}) \\ \vdots & \vdots & \vdots & \vdots \\ E(H_{k1}) & E(H_{k2}) & \dots & E(H_{kk}) \end{bmatrix}$$

- Let $H_{jj} = \frac{\partial^2 \mathcal{L}}{\partial \theta_j^2}$
- Let $H_{jk} = \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k}$
- The **Expected Fisher Information Matrix** is then the expected value of the Hessian.

$$I(\theta) = \begin{bmatrix} E(H_{11}) & E(H_{12}) & \dots & E(H_{1k}) \\ E(H_{21}) & E(H_{22}) & \dots & E(H_{2k}) \\ \vdots & \vdots & \vdots & \vdots \\ E(H_{k1}) & E(H_{k2}) & \dots & E(H_{kk}) \end{bmatrix}$$

- The **Jacobian** is the *inverse* of the EFI matrix

$$J(\theta) = I^{-1}(\theta)$$

- Yes ... the *inverse*. Brush off your notes.

Finding the multivariate MLE

1. Calculate the likelihood for your data.

Finding the multivariate MLE

1. Calculate the likelihood for your data.
2. Take the log.
3. Take the graident, set equal to $\mathbf{0}$, and solve.

Finding the multivariate MLE

1. Calculate the likelihood for your data.
2. Take the log.
3. Take the graident, set equal to $\mathbf{0}$, and solve.
4. Check your second order conditions if necessary.
5. Find the jacobian to get the variance covariance matrix.

Properties of the multivariate MLE

- Under appropriate regularity conditions:

$$(\hat{\theta} - \theta) \approx N(0, J_n)$$

- Also, if $\hat{\theta}_j$ is the j^{th} component of $\hat{\theta}$, then

$$\frac{(\hat{\theta}_j - \theta_j)}{\sqrt{Var(\hat{\theta}_j)}}$$

  converges in distribution to $N(0, 1)$, where $Var(\hat{\theta}_j)$ is the j^{th} diagonal element of J_n.
- Further, the covariance between element j and k of $\hat{\theta}$, $Cov(\hat{\theta}_j, \hat{\theta}_k) \approx J_{jk}$.

Example: Poisson regression

In poisson regression, we assume that y_i is Poisson with parameter λ_i, where

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{1i})$$

Example: Poisson regression

In poisson regression, we assume that y_i is Poisson with parameter λ_i, where

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{1i})$$

- Let $\beta = (\beta_0, \beta_1)'$
- $x_i = (1, x_{1i})'$

Example: Poisson regression

In poisson regression, we assume that y_i is Poisson with parameter λ_i, where

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{1i})$$

- Let $\beta = (\beta_0, \beta_1)'$
- $x_i = (1, x_{1i})'$
- Thus, $\lambda_i = \exp(x_i'\beta)$

1. Calculate the likelihood

$$L(\beta) \propto \prod_{i=1}^{N} \lambda_i^{y_i} e^{\lambda_i}$$

1. Calculate the likelihood

$$L(\beta) \propto \prod_{i=1}^{N} \lambda_i^{y_i} e^{\lambda_i}$$

$$\mathcal{L}(\beta) \propto \sum_{i=1}^{n} y_i x_i' \beta - \exp(x'\beta)$$

2. Take the derivative:

$$\nabla \mathcal{L}(\beta) = \left(\begin{array}{c} \sum_{i=1}^{n} (y_i - \exp(x'\beta))1 \\ \sum_{i=1}^{n} (y_i - \exp(x'\beta))x_{1i} \end{array} \right)$$

1. Calculate the likelihood

$$L(\beta) \propto \prod_{i=1}^{N} \lambda_i^{y_i} e^{\lambda_i}$$

$$\mathcal{L}(\beta) \propto \sum_{i=1}^{n} y_i x_i' \beta - \exp(x'\beta)$$

2. Take the derivative:

$$\nabla \mathcal{L}(\beta) = \begin{pmatrix} \sum_{i=1}^{n} (y_i - \exp(x'\beta))1 \\ \sum_{i=1}^{n} (y_i - \exp(x'\beta))x_{1i} \end{pmatrix}$$

Ok, great! Now we just need to solve for zero

The good, the bad, and the ugly of MLE

The good:

- MLE is sort of automatic.
- Tends to be unbiased, fairly efficient (in finite samples), and asymptotically efficient.
- When it works it is wicked fast.

The bad:

- Closed form solutions are rarely available for interesting problems
- So we often have to estimate the MLE numerically

The ugly:

- Numerical methods can often suck and do not produce errors to tell you that.
- For huge classes of problems, additional "tricks" must be adopted to find MLE that move us away from some of the nice properties discussed above.

Next class

- We will work through some numerical approaches to finding the MLE
- Newton-Raphson, IWLS, GMM, and gradient descent

Side note on Fisher information

Let $X_i, \ldots, X_n$ be iid variables from a distribution in the exponential family (which ensures some regularity conditions and that the log-likelihood is twice differentiable).

We want to show that

$$E\left[\left(\frac{\partial}{\partial\theta}\mathcal{L}(\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta)\right]$$

$$\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta) = \frac{\frac{\partial^2}{\partial\theta^2}L(\theta)}{L(\theta)} - \left(\frac{\frac{\partial^2}{\partial\theta^2}L(\theta)}{L(\theta)}\right)^2$$

$$= \frac{\frac{\partial^2}{\partial\theta^2}L(\theta)}{L(\theta)} - \left(\frac{\partial}{\partial\theta}\mathcal{L}(\theta)\right)^2$$

Now, in expectation the second derivative of $E(L(\theta)) = E(f(x|\theta)) = \int f(x|\theta)dx$ is going to be zero (assertion, but it makes sense for all of the examples we have looked at).

So, since the negative can come out of the expectation, we now have that

$$E\left[\left(\frac{\partial}{\partial\theta}\mathcal{L}(\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta)\right]$$