GOV 2001/ 1002/ Stat E-200 Section 5
Maximum Likelihood Estimation

Solé Prillaman

Harvard University

February 25, 2015

Logistics

Paper Selection- Comments coming soon!

Reading Assignment- Franklin (1991) and UPM Ch 4.

Problem Set 4- Due by 6pm next Wednesday on Canvas.

Assessment Question- Due by 6pm next Wednesday on Canvas. You must work alone and only one attempt.

Replication Paper

Start getting data- journal website, author's website, ICPSR, email author.

Start replicating the paper.

Replication- due March 25 by 6pm.

Outline

Maximum Likelihood Estimators

Steps to finding the MLE:

1. Write out the model.
2. Calculate the likelihood ($\mathrm{L}(\theta|y)$) for all observations.
3. Take the log of the likelihood ($\ell(\theta|\mathbf{Y})$).
4. Plug in the systematic component for θ_i.
5. Bring in observed data.
6. Maximize $\ell(\theta|y)$ with respect to θ and confirm that this is a maximum.
7. Find the variance of your estimate.

Univariate Example

<u>Ex</u>. Waiting for the Redline – How long will it take for the next T to get here?

1. Write the Model

Y is a Exponential random variable with parameter λ.

$$f(y) = \lambda e^{-\lambda y}$$

Let's assume that Y is distributed Exponentially with some constant rate of seeing a train arrive across observations. We will also assume that observations are independent (this is an important assumption).

The model:

1. $Y_i \sim f_{\text{expo}}(y_i|\lambda_i)$.
2. $\lambda_i = \lambda$.
3. Y_i and Y_j are independent for all $i \neq j$.

2. Calculate $L(\lambda|y)$

Remember that with one observation $L(\lambda|y) \propto p(y|\lambda)$.

$$\begin{aligned}
L(\lambda_i|y_i) &\propto f_{\text{expo}}(y_i|\lambda_i) \\
&\propto \lambda_i e^{-\lambda_i y_i} \\
L(\lambda|y) &\propto \lambda_1 e^{-\lambda_1 y_1} \cdot \lambda_2 e^{-\lambda_2 y_2} \cdot \ldots \cdot \lambda_n e^{-\lambda_n y_n} \\
&\propto \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i}
\end{aligned}$$

3. Calculate $\ell(\lambda|y)$

$$
\begin{aligned}
L(\lambda|y) &\propto \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i} \\
\ell(\lambda|y) &= \ln\left(\prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i}\right) + C \\
&= \sum_{i=1}^{n} \ln(\lambda_i e^{-\lambda_i y_i}) + C \\
&= \sum_{i=1}^{n} (\ln(\lambda_i) + \ln(e^{-\lambda_i y_i})) + C \\
&= \sum_{i=1}^{n} (\ln \lambda_i - \lambda_i y_i) + C
\end{aligned}
$$

4. Plug in Systematic Component

Remember in our model we assumed that the rate was constant across observations (i.e. $\lambda_i = \lambda$).

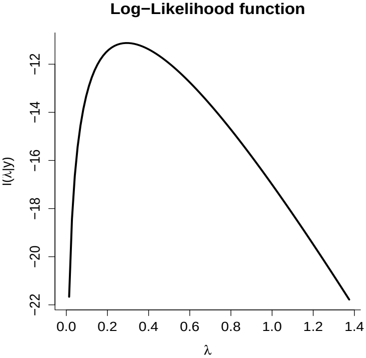
$$\begin{aligned}
\ell(\lambda|y) &= \sum_{i=1}^{n}(\ln\lambda_i - \lambda_i y_i) + C \\
&= \sum_{i=1}^{n}(\ln\lambda - \lambda y_i) + C \\
&= n\ln\lambda - \lambda\sum_{i=1}^{n} y_i + C
\end{aligned}$$

5. Bring in Observed Data

Let's say that we take the train 5 times this week and observe the following times until arrival $Y : \{1, 5, 8, 2, 1\}$.

$$
\begin{aligned}
\ell(\lambda|y) &= n \ln \lambda - \lambda \sum_{i=1}^{n} y_i + C \\
&= 5 \ln \lambda - \lambda(1 + 5 + 8 + 2 + 1) + C \\
&= 5 \ln \lambda - 17\lambda + C
\end{aligned}
$$

6. Maximize $\ell(\theta|Y)$



Where is the maximum?
How could we find it analytically using

$$\ell(\lambda|y) = 5\ln\lambda - 17\lambda + C?$$

Optimization

Optimization is the process of minimizing or maximizing a function by systematically choosing the values of variables from within an allowed set. For example:

$$\min_{x \in [-\infty, \infty]} f(x) = -\frac{1}{2}(3 - x)^2$$

- $f(x)$ is called the objective function
- x is the parameter (for us λ or β, π, σ, etc.)
- $x \in [-\infty, \infty]$ is the allowed set or the parameter space

Two ways to solve:

1. Analytically
2. Numerically

Analytic Optimization

Step One: Take the first derivative of the function with respect to the parameter of interest.

- The derivative of a function at a value x_0, denoted by $f'(x_0)$ or $\frac{\partial f}{\partial x}(x_0)$, is the instantaneous rate of change in $f(x)$ at x_0.
- Define the **Score** as:

$$S(\theta) = \frac{\partial \ell(\theta|\mathbf{x})}{\partial \theta}$$

Analytic Optimization

Step One: Take the first derivative of the function with respect to the parameter of interest.

$$
\begin{aligned}
\ell(\lambda|y) &= 5\ln\lambda - 17\lambda + C \\
S(\lambda) = \frac{\partial \ell(\lambda|y)}{\partial\lambda} &= \frac{\partial(5\ln\lambda - 17\lambda + C)}{\partial\lambda} \\
&= \frac{5}{\lambda} - 17
\end{aligned}
$$

Analytic Optimization

Step Two: Set the first derivative of the function equal to 0 and identify the critical value(s) of our parameter.

- $f'(x_0)$ describes the behavior of a function on an interval [a,b]
  - If $f^{'}(x) > 0$ for all $x \in [a, b]$, then f is increasing on the interval $[a, b]$
  - If $f^{'}(x) < 0$ for all $x \in [a, b]$, then f is decreasing on the interval $[a, b]$
  - If $f^{'}(x) = 0$ at some $x \in [a, b]$ then we say x is a critical value of f.

Analytic Optimization

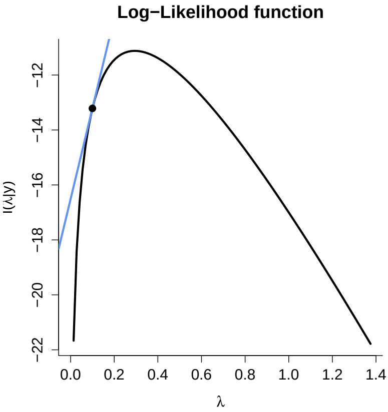
Step Two: Set the first derivative of the function equal to 0 and identify the critical value(s) of our parameter.

$$\begin{aligned}\frac{\partial \ell(\lambda|y)}{\partial \lambda} &= 0\\ \frac{5}{\lambda} - 17 &= 0\\ \hat{\lambda} &= \frac{5}{17}\end{aligned}$$

$\hat{\lambda}$ is the maximum likelihood estimate given our observed data!

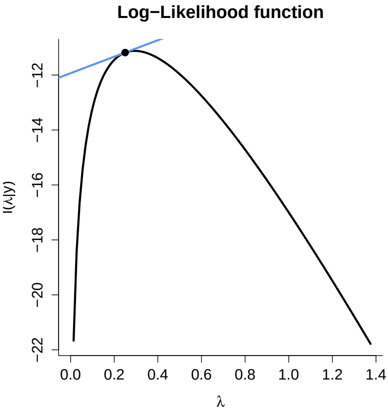
Analytic Optimization

Why do we think $\hat{\lambda}$ is the MLE? Put another way, why do we think $\hat{\lambda}$ maximizes the log-likelihood function?



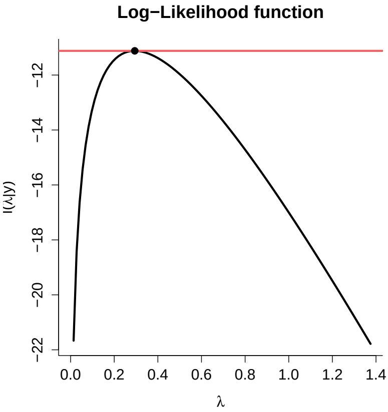
Analytic Optimization

Why do we think $\hat{\lambda}$ is the MLE? Put another way, why do we think $\hat{\lambda}$ maximizes the log-likelihood function?



Analytic Optimization

Why do we think $\hat{\lambda}$ is the MLE? Put another way, why do we think $\hat{\lambda}$ maximizes the log-likelihood function?



Analytic Optimization

Step Three: Compute the second derivative of the function at the critical value(s) and evaluate.

- The second derivative of a function $f^{''}(x)$ or $\frac{\partial^2 f}{\partial x \partial x}(x)$ is the derivative of the derivative, or the rate of change of the rate of change.
- Use the following to evaluate your critical value(s):
  - If $f'(x_0) = 0$, and $f^{''}(x_0) < 0$, then x_0 is a maximum
  - If $f'(x_0) = 0$ and $f^{''}(x_0) > 0$, then x_0 is a minimum
  - If $f'(x_0) = 0$ and $f^{''}(x_0) = 0$, then x_0 may be a minimum, a maximum, or neither.
- Define the **Hessian** for the univariate case as:

$$H(\theta) = \frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}$$

Analytic Optimization

Step Three: Compute the second derivative of the function at the critical value(s) and evaluate.

$$\begin{aligned} S(\lambda) = \frac{\partial \ell(\lambda|y)}{\partial \lambda} &= \frac{5}{\lambda} - 17 \\ H(\lambda) = \frac{\partial^2 \ell(\lambda|y)}{\partial \lambda^2} &= -\frac{5}{\lambda^2} \\ \frac{\partial^2 \ell(\lambda|y)}{\partial \lambda^2}\Big|_{\lambda=\hat{\lambda}} &= -\frac{5}{(5/17)^2} \\ &= -\frac{289}{5} \end{aligned}$$

This confirms that $\hat{\lambda}$ is a **maximum** because the second derivative is negative.

Numerical Optimization

For problems of only slightly more complexity, using derivatives and solving for parameters in order to maximize may be not only impractical but impossible.

There are a number of functions in R which can be used to optimize functions, but the one we will use most heavily is `optim()`.

`optim()` takes a starting value (`par`) and a function (`fn`) as its main arguments.

Numerical Optimization

Step One: Write a function in R for your log-likelihood.

```
ll.expo <- function(lambda, y){
  length(y)*log(lambda) - lambda*sum(y)
}
```

`par` is our parameter lambda.

NUMERICAL OPTIMIZATION

Step Two: Optimize this function using `optim()`.

```
# Create our data
y <- c(1,5,8,2,1)

# Optimize
opt.expo <- optim(par = 0.01, fn = ll.expo, y = y,
    method = "BFGS", control = list(fnscale = -1),
    hessian = TRUE)
```

What are these three extra arguments?

1. `method` is the algorithm used to find the maximum.
2. `fnscale` multiplies the function by the given constant. As a default `optim()` finds the minimum so multiplying our function by -1 fools `optim()` into finding the maximum.
3. `hessian = TRUE` requests that `optim` return a matrix of second derivatives which in the univariate case will be 1x1.

Numerical Optimization

```
## Output from Optim
opt.expo

$par
[1] 0.2941228

$value
[1] -11.11888
...
$hessian
          [,1]
[1,] -57.79931

# If we want to pull out our MLE
mle <- opt.expo$par
# This is the same as 5/17, our analytic MLE

# If we want to pull out the matrix of second derivatives evaluated
hessian <- opt.expo$hessian
# This is the same as -289/5, our analytic hessian
```

WHAT IS `OPTIM` DOING?

It depends on your choice in the `method` argument.

- `Nelder-Mead`: this is the default; it is slow but somewhat robust to non-differentiable functions.
- `BFGS`: a quasi-Newton Method; it is fast but needs a well behaved objective function.
- `L-BFGS-B`: similar to BFGS but allows box-constraints (i.e. upper and lower bounds on variables).
- `CG`: conjugate gradient method, may work for really large problems (we won't really use this).
- `SANN`: uses simulated annealing – a stochastic global optimization method; it is very robust but *very* slow.

7. Find the Variance of $\hat{\lambda}$

We are interested in calculating a measure of uncertainty of our MLE. That is, we are after the following:

$$\mathrm{Var}(\hat{\theta}_{MLE})$$

Conceptually, we want to know how much information the MLE contains about the underlying parameter.

Asymptotic Distribution of the MLE

It can be shown by the **Central Limit Theorem** that under certain regularity conditions, the MLE is distributed normally with a mean equal to the true parameter (θ_0) and the variance equal to the inverse of the expected sample Fisher information at the true parameter (denoted as $\mathcal{I}_n(\theta_0)$):

$$\hat{\theta}_{MLE} \sim \mathcal{N}\left(\theta, \Big(\underbrace{-E\left[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2}\Big|_{\theta=\theta}\right]}_{\mathcal{I}_n(\theta)}\Big)^{-1}\right)$$

Asymptotic Distribution of the MLE

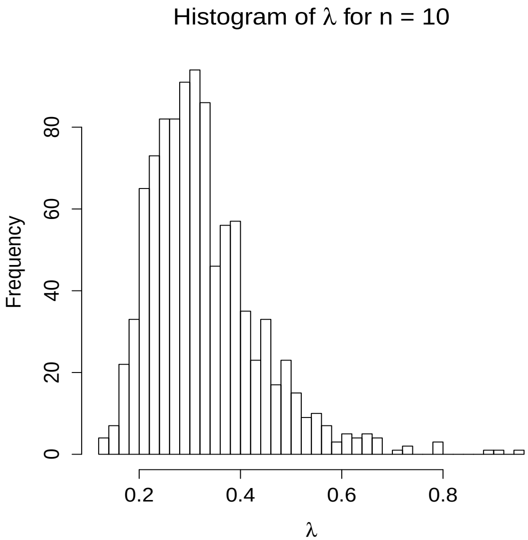
Let's convince ourselves of this. We're going to generate 1000 datasets, of 10 observations each from an Exponential with $\lambda = .5$.

```
n <- 10
data <- sapply(seq(1,1000),function(x)
      rexp(n, rate=.5))
dim(data)
[1]   10 1000
```

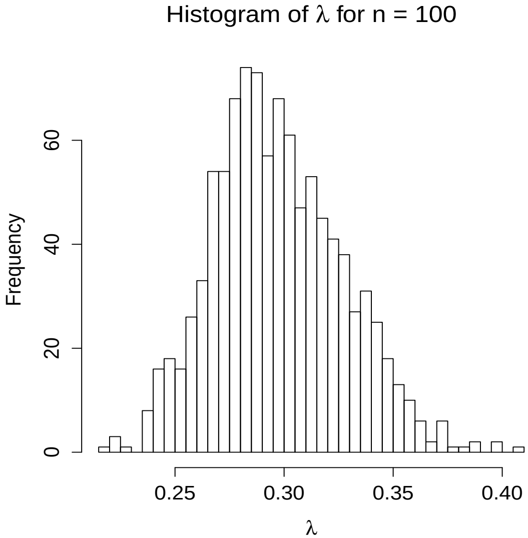
For each of these datasets, we're going to find the maximum likelihood estimate for λ.

```
llexp <- function(param, y){sum(dexp(y, rate=param, log=T))}
out <- NULL
for(i in 1:1000){
out[i] <- optim(c(1), fn=llexp, y=data[,i],
   method="BFGS", control=list(fnscale=-1))$par
}
```

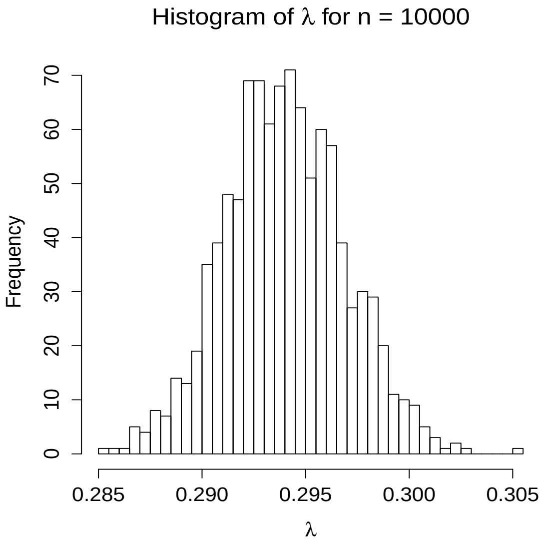
Asymptotic Distribution of the MLE



Asymptotic Distribution of the MLE



Asymptotic Distribution of the MLE



Asymptotic Distribution of the MLE

How do we think about this intuitively?

- The Central Limit Theorem states that the mean of independent random variables will become approximately Normal as n goes to infinity.
- But we're not talking about the mean!
- Yes, but the maximum of the log-likelihood is essentially the mean of a lot of likelihoods. And we use this maximum to estimate our parameter.
- Therefore, as n gets larger, and the more likelihoods we have to conglomerate, the more normal the distribution of the parameter becomes.

Asymptotic Distribution of the MLE

So for large n, our parameter θ is distributed Normally with the mean as the true value of θ and variance $[\mathcal{I}(\theta)]^{-1}$:

$$\hat{\theta}_{MLE} \sim \mathcal{N}\left(\theta, \left(\mathcal{I}(\theta)\right)^{-1}\right)$$

Variance of MLE

- Measure of curvature: Hessian Matrix

$$H(\hat{\theta}) = \frac{\partial^2 \ell(\theta)}{\partial^2 \theta}\bigg|_{\theta=\hat{\theta}}$$

- Measure of Information: Fisher Information Matrix

$$I(\hat{\theta}) = -H(\hat{\theta}) = -\frac{\partial^2 \ell(\theta)}{\partial^2 \theta}\bigg|_{\theta=\hat{\theta}}$$

- Estimate of Variance: Inverse of the Fisher Information Matrix

$$\hat{Var}(\hat{\theta}) \approx [I(\hat{\theta})]^{-1} \approx \left[-\frac{\partial^2 \ell(\theta)}{\partial^2 \theta}\bigg|_{\theta=\hat{\theta}} \right]^{-1}$$

- Estimate of Standard Error: Square root of $Var(\hat{\theta})$

$$\hat{SE}(\hat{\theta}) = \sqrt{\hat{Var}(\hat{\theta})} = \sqrt{\left[-\frac{\partial^2 \ell(\theta)}{\partial^2 \theta}\bigg|_{\theta=\hat{\theta}} \right]^{-1}}$$

Variance of MLE

- Asymptotically:

$$\text{Var}(\hat{\theta}_{MLE}) = [\mathcal{I}_n(\theta)]^{-1} = \left(-E\left[\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial \theta^2} \right] \right)^{-1}$$

- That is, it's the inverse of the **Expected Fisher Information** evaluated at the true parameter θ
- Conceptually, this is the *expected* curvature of the log-likelihood curve across repeated samples at the point θ
- We estimate this with the inverse of **Observed Fisher Information**:

$$\hat{\text{Var}}(\hat{\theta}_{MLE}) = [I(\hat{\theta})]^{-1} = \left(- \frac{\partial^2 \ell(\theta)}{\partial^2 \theta} \bigg|_{\theta=\hat{\theta}} \right)^{-1}$$

- As $n \to \infty$, the observed Fisher information converges to the expected Fisher information and the $\hat{\theta}_{MLE}$ converges to θ_0

7. Find the Variance of $\hat{\lambda}$

Remember that we already calculated the Hessian when we checked that $\hat{\lambda}$ is a maximum:

$$H(\hat{\lambda}) = -\frac{289}{5}$$

We can use this to calculate the standard error of $\hat{\lambda}$.

$$\begin{aligned} I(\hat{\lambda}) &= \frac{289}{5} \\ \hat{Var}(\hat{\lambda}) &= \frac{5}{289} \\ \hat{SE}(\hat{\lambda}) &= \sqrt{\frac{5}{289}} \end{aligned}$$

Our Univariate Example

Given our model and data

$$
\begin{aligned}
Y_i &\sim f_{\text{expo}}(y_i|\lambda_i) \\
\lambda_i &= \lambda \\
Y &: \{1, 5, 8, 2, 1\}
\end{aligned}
$$

- $L(\lambda|y) \propto \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i}$
- $\ell(\lambda|y) = 5 \ln \lambda - 17\lambda + C$
- $S(\lambda) = \frac{5}{\lambda} - 17$
- $\hat{\lambda} = \frac{5}{17}$
- $H(\lambda) = -\frac{5}{\lambda^2}$
- $\hat{SE}(\hat{\lambda}) = \sqrt{\frac{5}{289}}$

Outline

Multivariate

Ex. Waiting for the Redline – How long will it take for the next T to get here?

But this time we want to add covariates. What do you think affects the wait for the Redline?

1. Write the Model

How would we model this?

- We know the stochastic component:

$$Y_i \sim f_{\text{expo}}(y_i|\lambda_i)$$

- Remember, for an Exponential

$$\mu_i = \frac{1}{\lambda_i}$$

- So we're going to set the systematic component

$$\begin{aligned} \mu_i &= \exp(X_i\beta) \\ \lambda_i &= \frac{1}{\exp(X_i\beta)} \end{aligned}$$

What are the parameters?

2. Calculate $L(\lambda|y)$ and 3. $\ell(\lambda|y)$

$$
\begin{aligned}
L(\lambda|y) &= \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i} \\
\ell(\lambda|y) &= \sum_{i=1}^{n} (\ln \lambda_i - \lambda_i y_i) + C
\end{aligned}
$$

4. Plug in Systematic Component

$$
\begin{aligned}
\ell(\beta|y) &= \sum_{i=1}^{n}(\ln\lambda_i - \lambda_i y_i) + C \\
&= \sum_{i=1}^{n}\left(\ln\frac{1}{exp(X_i\beta)} - \frac{1}{exp(X_i\beta)}y_i\right) \\
&= \sum_{i=1}^{n}\left(\ln 1 - \ln(exp(X_i\beta)) - \frac{1}{exp(X_i\beta)}y_i\right) \\
&= \sum_{i=1}^{n}\left(-(X_i\beta) - \frac{1}{exp(X_i\beta)}y_i\right)
\end{aligned}
$$

5. Bring in Observed Data

I'm going to say whether or not it is Friday and the minutes behind schedule are important covariates.

Let's create some fake data:

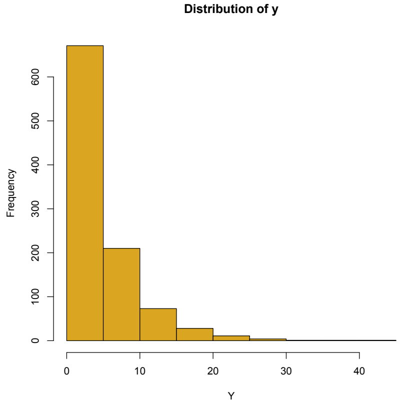
```
set.seed(02139)
n <- 1000
Friday <- sample(c(0,1), n, replace=T)
minsSch <- rnorm(n, 3, .5)

Y <- rexp(n, rate = 1/exp(1.25 - .5*Friday +.2*minsSch))
data <- as.data.frame(cbind(Y, Friday, minsSch))
```

5. Bring in Observed Data

Let's look at Y:

```
hist(Y, col = "goldenrod", main = "Distribution of y")
```



6. Maximize $\ell(\beta|y)$

Remember the log-likelihood we solved for before:

$$\ell(\beta|y) \quad = \quad \sum_{i=1}^{n} \left(-(X_i\beta) - \frac{1}{\exp(X_i\beta)} y_i \right)$$

We can find the MLE by setting the score to 0.

- Score:

$$S(\boldsymbol{\beta}) = \nabla \ell(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_k} \end{pmatrix}$$

Set **each** element to 0 and solve the system of equations to get $\hat{\beta}$.

6. MAXIMIZE $\ell(\beta|y)$

OR, we can solve for the MLE in R by first programming the log-likelihood.

```
llexp <- function(param, y, x){
rate <- 1/exp(x%*%param)
sum(dexp(y, rate=rate, log=T))
}

llexp2 <- function(param, y,x){
cov <- x%*%param
sum(-cov - 1/exp(cov)*y)
}
```

6. MAXIMIZE $\ell(\beta|y)$

We can maximize our function using `optim`:

```
#Create X with an intercept
X <- cbind(1, Friday, minsSch)

#Specify starting values for all three parameters
param <- c(1,1,1)

#Solve using optim
out <- optim(param, fn=llexp, y=Y, x=X, method="BFGS",
   hessian=T, control=list(fnscale=-1))

out$par
[1]  1.0885871 -0.4634621  0.2120591
```

7. Find the Variance of $\hat{\beta}$

We can find the variance of the MLE with:

$$\hat{Var}(\hat{\beta}) = [-H(\hat{\beta})]^{-1}$$

But with multiple parameters, the Hessian is now:

- Hessian:

$$H(\hat{\beta}) = \begin{pmatrix} \frac{\partial^2}{\partial \beta_1^2} & \frac{\partial^2}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2}{\partial \beta_1 \partial \beta_k} \\ \frac{\partial^2}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2}{\partial \beta_2^2} & \cdots & \frac{\partial^2}{\partial \beta_2 \partial \beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \beta_k \partial \beta_1} & \frac{\partial^2}{\partial \beta_k \partial \beta_2} & \cdots & \frac{\partial^2}{\partial \beta_k^2} \end{pmatrix} \Bigg|_{\beta=\hat{\beta}}$$

7. Find the Variance of $\hat{\beta}$

We can calculate this in R using the output from `optim`.

```
# Get the Hessian from optim
H <- out$hessian

# Calculate the observed fisher information
I <- -H

# Calculate the variance-covariance matrix
V <- solve(I)

# Get the standard errors
ses <- sqrt(diag(V))
```

Outline

Likelihood Ratio Test

- Useful for when you are comparing two models.
- We'll call these restricted and unrestricted:

$$\begin{aligned} \textit{Unrestricted} &: \beta_0 + \beta_1 X_1 + \beta 2 X_2 \\ \textit{Restricted} &: \beta_0 + \beta_2 X_2 \end{aligned}$$

- We want to test the usefulness of the parameters in the unrestricted model but omitted in the restricted model (eg. β_1).

Likelihood Ratio Tests

Here's how to operationalize this:

- Let L_u^* be the maximum of the unrestricted likelihood, and let L_r^* the maximum of the restricted likelihood.
- But adding more variables can only increase the likelihood.
- Thus, $L_u^* \geq L_r^*$, or $\frac{L_r^*}{L_u^*} \leq 1$ always.
- If the likelihood ratio is exactly 1, then there's no effect of the extra parameters at all ($L_u^* = L_r^*$).

Likelihood Ratio Test

Now, let's define a test statistic:

$$\begin{aligned}
\text{define } : \Re &= -2\ln\frac{L_r^*}{L_u^*} \\
&= 2(\ln L_u^* - \ln L_r^*) \\
\Re &\sim \chi^2_m \\
\text{Reject if } : \Re &> \chi^2_{m,1-\alpha}
\end{aligned}$$

- m is the number of restrictions.
- $\Re$ will always be greater than zero.
- Key question: how much greater than zero does $\Re$ have to be in order to convince us that the difference is due to systematic differences between the two models?

Back to Our Example

What if we wanted to test whether the minutes behind schedule should be in our model at all?

```
unrestricted <- optim(param, fn=llexp, y=Y, x=X,
      method="BFGS", hessian=T, control=list(fnscale=-1))
unrestricted$value
[1] -2503.445
```

v.

```
restricted <- optim(c(1,1), fn=llexp, y=Y,
      x=cbind(1, Friday), method="BFGS",
      hessian=T, control=list(fnscale=-1))
restricted$value
[1] -2509.471
```

Back to Our Example

Under the null that the restrictions are valid, the test statistic would be distributed χ^2 with one degree of freedom:

```
# Calculate our test statistic
r <- 2*(unrestricted$value - restricted$value)

# Calculate the p-value for this test statistic
1-pchisq(r,df=1)
[1] 0.0005176814
```

So the probability of getting this test statistic under the null is extremely small. We reject.

QUESTIONS

Questions?

Outline

Optimization Strategies

This appendix provides alternative methods of optimization.

Newton's Method

Newton's method: a pretty good approach for a continuous and twice-differentiable function. We'll look at a univariate function here. Suppose we know our function $f(\cdot)$ and we have

a starting value of x_0. Our goal is to find move from x_0 to x_1 such that $f'(x_1) = 0$ (or is at least closer to 0 then than at x_0). This will be a sequential process of approximation and

eventually $f'(x_n)$ will be close enough to zero to let us declare that x_n a critical value.

Taylor Series Expansion: A Staple of Calculus

We can approximate function $f(\cdot)$ at point a using a Taylor series expansion:

$$\boxed{f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \ldots}$$

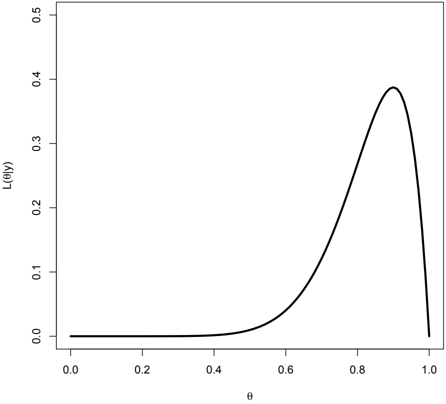
In this course, we'll work with the first- and second-order Taylor polynomials:

$$f(x) \approx f(a) + f'(a)(x-a)$$

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2$$

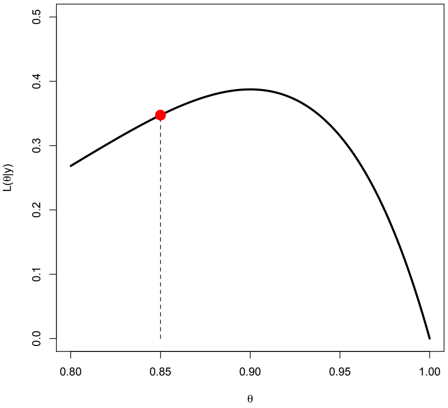
Likelihood of Binomial Distribution

This is the likelihood of a binomial distribution with 10 trials ($N = 10$) from which we drew one observation: $y = 9$.



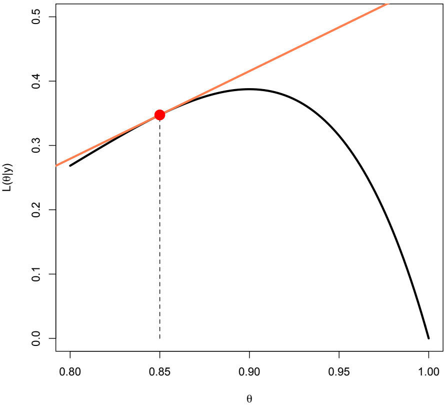
Likelihood of Binomial Distribution: Zoomed In

We want to approximate the likelihood curve around $\theta_0 = 0.85$.



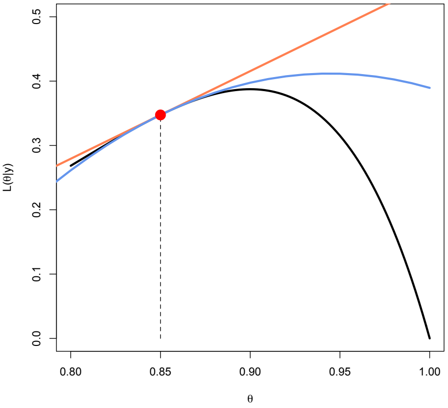
Likelihood of Binomial Distribution: Zoomed In

We can approximate the likelihood around $\theta_0 = 0.85$ using a first-order Taylor polynomial.



Likelihood of Binomial Distribution: Zoomed In

We can improve our approximation of the likelihood around $\theta_0 = 0.85$ by using a second-order Taylor polynomial.



Why Don't We Just Maximize the Second–Order Taylor Polynomial?

We can write a second-order Taylor expansion around θ_0 as:

$$f(\theta) \approx f(\theta_0) + f'(\theta_0)(\theta - \theta_0) + \frac{f''(\theta_0)}{2!}(\theta - \theta_0)^2$$

To maximize, take the derivative with respect to θ and set it equal to 0:

$$f'(\theta) = f'(\theta_0) + f''(\theta_0)(\theta - \theta_0) = 0$$

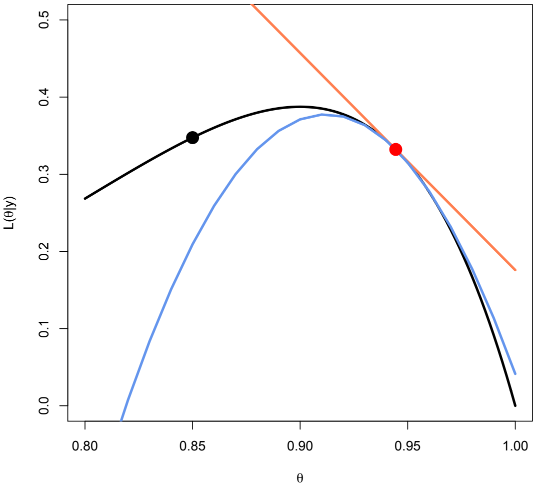
Rearranging:

$$\theta = \theta_0 - \frac{f'(\theta_0)}{f''(\theta_0)}$$

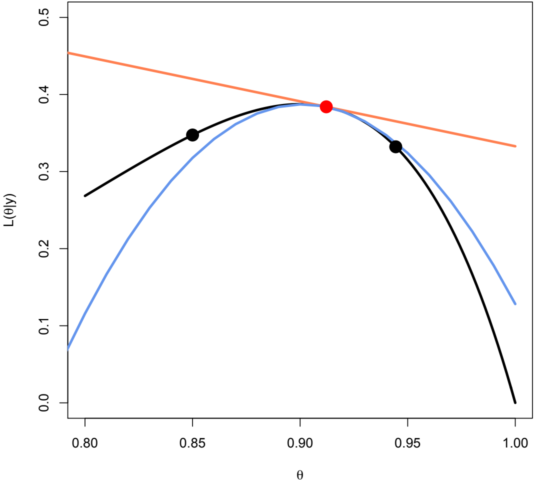
We've Just Derived the Update Step We Wanted!

$$\boxed{\theta_{n+1} = \theta_n - \frac{f'(\theta_n)}{f''(\theta_n)}}$$

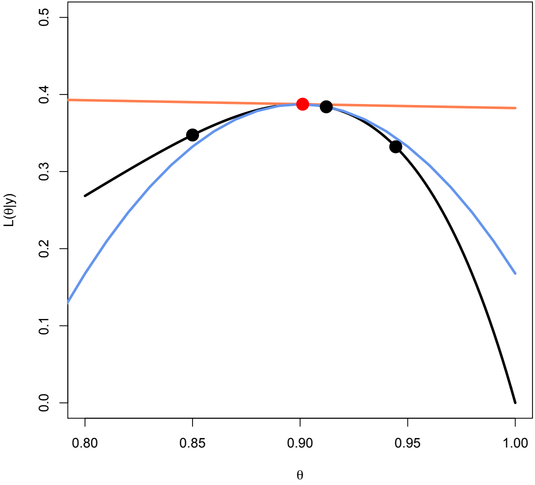
FIRST ITERATION



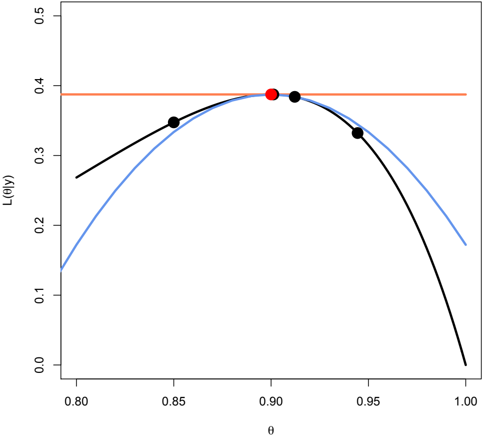
Second Iteration



Third Iteration



Fourth Iteration



Another Example of Taylor Series Approximation

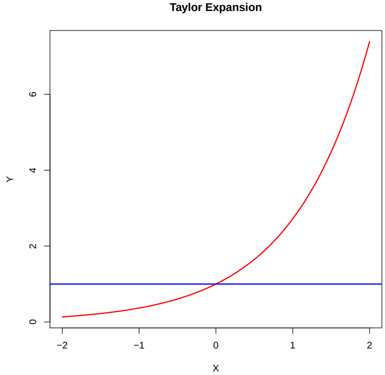


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_0(x_1) = 1$ (from Wikipedia)

Another Example of Taylor Series Approximation

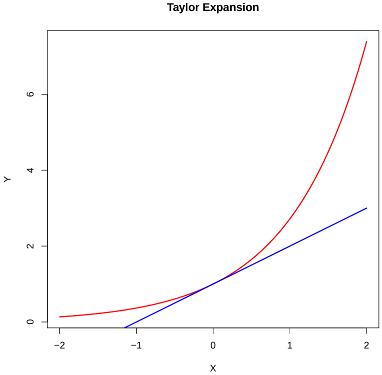


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_1(x_1) = 1 + x_1$ (from Wikipedia)

Another Example of Taylor Series Approximation

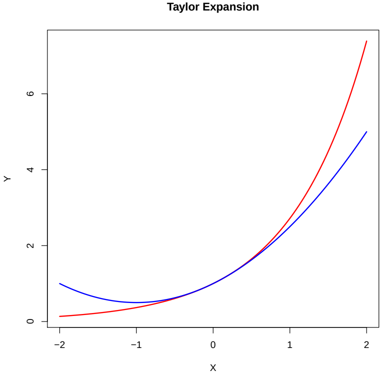


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0, g_2(x_1) = 1 + x_1 + \frac{x^2}{2}$ (from Wikipedia)

Another Example of Taylor Series Approximation

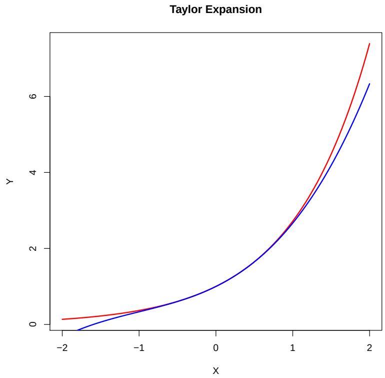


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0, g_3(x_1) = 1 + x_1 + \frac{x^2}{2} + \frac{x^3}{6}$ (from Wikipedia)

Another Example of Taylor Series Approximation

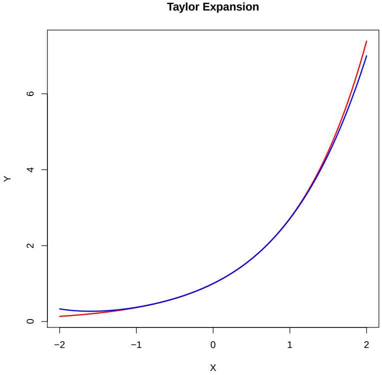


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_4(x_1) = 1 + x_1 + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}$ (from Wikipedia)

Newton in Action: Bernoulli Example Revisited

Let's maximize our likelihood: $3\ln\pi + 2\ln(1-\pi)$.

Recall that $L'(\pi) = \frac{3}{\pi} - \frac{2}{1-\pi}$ and $L''(\pi) = -\frac{3}{\pi^2} - \frac{2}{(1-\pi)^2}$.

Starting at $\pi_0 = .3$, we use our updating formula:

$$\pi_1 = \pi_0 - \frac{L'(\pi_0)}{L''(\pi_0)} = .3 - \frac{L'(.3)}{L''(.3)} = 0.4909.$$

Now use $\pi_1 = .4909$ as a starting value.

$$\pi_2 = \pi_1 - \frac{L'(\pi_1)}{L''(\pi_1)} = .4909 - \frac{L'(.4909)}{L''(.4909)} = 0.5991.$$

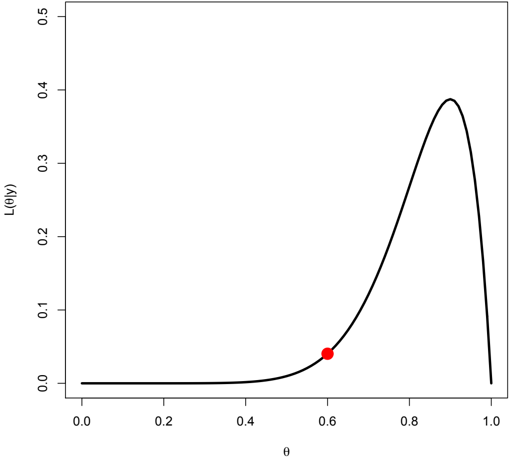
So we're already there!

Properties of Newton-Raphson

- Converges quickly
- Can get stuck in local minima/maxima
- Can have troubles with root jumping
- Won't walk at all on a flat space

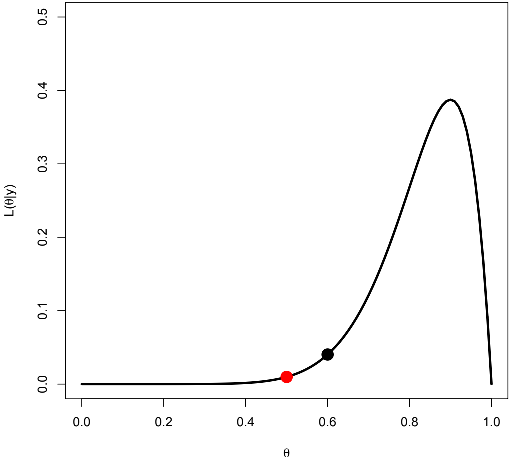
Newton-Raphson Gone Awry

What if we had taken $\theta_0 = 0.60$ to be the starting point for the Newton-Raphson maximization for the binomial likelihood?



It Walks the Wrong Way...

The first iteration:



And Then Gets Stuck!

After 10 iterations: