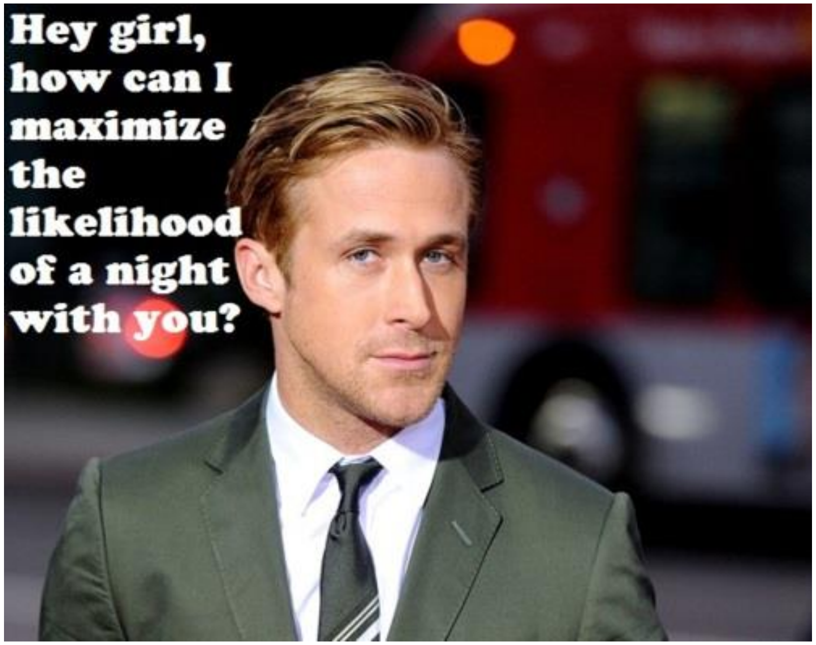


BIOS 6612
Lecture 3

Logistic Regression II
Maximum Likelihood Estimation

KKMN Chapters 20,21
Vittinghoff. Regression Methods in Biostatistics. Chapter 6
Agretsi (2002) Catgorical Data Analysis, 2nd Edition. Section 4.2, Chapter 5 up to 5.1.3, Section 6.6

Review (Lecture 2) / Current (Lecture 3)/ Preview (Lecture 4)

- Lecture 2: Introduction of Logistic Regression
  - Introduction to logistic regression
    - $\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} \Rightarrow 0 < p < 1$

- Lecture 3: Maximum Likelihood Estimation
  - Maximum Likelihood Estimation (MLE)
    - Analytic solution for intercept only model

- Lecture 4: Wald, Score, & Likelihood Ratio Tests
  - When to use each
  - Issues with Wald

Maximum Likelihood Estimation (MLE)

- The MLE of θ is the value, $\hat{\theta}$ which maximizes the likelihood $L(\theta)$ or the log –likelihood $\log\ L(\theta)$
  - The value of $\hat{\theta}$ that maximizes the likelihood also maximizes the log-likelihood since the log-likelihood is a monotone function of the likelihood
  - Usually easier to maximize the log-likelihood, log L(θ)
- Want to solve $\frac{\partial \log\ L(\theta)}{\partial \theta} = 0$
- Technically, need to verify that it is a maximum rather than a minimum
  - i.e. $\left[\frac{\partial^2 \log\ L(\theta)}{\partial \theta^2}\right]_{\theta=\hat{\theta}} < 0$
- The negative of the second derivative is called the information
  - $\frac{-\partial^2 \log\ L(\theta)}{\partial \theta^2}$
    - Plays an important role in likelihood theory

Maximum Likelihood Estimation (Bernoulli Example)

Simple Case: Bernoulli (Y_i = 0 or 1)

- Suppose in a population from which we are sampling, each individual has the **same** probability, p, that an event occurs
  - where an event can be a disease, trait, etc

- We want to estimate p=P(Y=1), from a random sample of n individuals

- For each individual in our sample of size n,
  - $Y_i = 1$ indicates that an event occurs for the ith subject
  - otherwise, $Y_i = 0$

- Recall The probability mass function (p.m.f.) of Y can be written as

$$P(Y=y|p) = p^{y}(1-p)^{1-y},\ \ y = 0,1\ \ ,\ 0 \le p \le 1$$

$$P(Y=1) = p^{1}(1-p)^{1-1} = p$$
$$P(Y=0) = p^{0}(1-p)^{1-0} = 1-p$$

$$P(Y=1)+ P(Y=0)=p+1-p=1$$

$$E(Y)= 0*P(Y=0)+1*P(y=1)=p$$

$$E(Y^2)= 0^2*P(Y=0)+1^2*P(y=1)=p$$
$$Var(Y)= E(Y^2)- E(Y)^2=p-p^2=p(1-p)$$

Maximum Likelihood Estimation (Bernoulli Example)

- The probability mass function (p.m.f.) of Y can be written as

$$P(Y=y|p) = p^{y}(1-p)^{1-y},\ y = 0,1\ \ , 0 \le p \le 1$$

- The observed data are $Y_1, \ldots, Y_n$.

- The joint probability of the data (the likelihood of the data) is a function of p, and is given by

$$L = \prod_{i=1}^{n} p^{Yi}(1-p)^{1-Yi} = p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i}$$

this is the probability that $Y_1=y_1$, $Y_2=y_2, \ldots, Y_n=y_n$.

- p does not depend on subject i
  - Intercept only model
  - Same probability for all subjects

Intercept only model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0$$

$$p_i = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

Maximum Likelihood Estimation (Bernoulli Example)

- For estimation, it is easier work with the log-likelihood

$$\text{Log(Likelihood)}:= LL = \ln(L) = \ln\left(\prod_{i=1}^{n} p^{Yi}(1-p)^{1-Yi}\right)$$

$$= \sum_{i=1}^{n} \ln\left(p^{Yi}(1-p)^{1-Yi}\right)$$

$$= \sum_{i=1}^{n} \left(\ln\left(p^{Yi}\right) + \ln\left((1-p)^{1-Yi}\right)\right)$$

$$= \sum_{i=1}^{n} \left(Y_i \ln(p) + (1-Y_i)\ln(1-p)\right)$$

$$= \underbrace{\left(\sum_{i=1}^{n} Y_i\right)}_{\text{\# obs with event}} \underbrace{\ln(p)}_{\text{probability of event}} + \underbrace{\left(n - \sum_{i=1}^{n} Y_i\right)}_{\text{\# obs without event}} \underbrace{\ln(1-p)}_{\text{probability no event}}$$

- The maximum likelihood estimate (MLE) of p is that value that maximizes LL (equivalent to maximizing L), which can be obtained numerically, or by setting the first derivative equal to 0.

Solving for the MLE of *p* (Bernoulli Example)

$$LL = \left(\sum_{i=1}^{n} Y_i\right)\ln(p) + \left(n - \sum_{i=1}^{n} Y_i\right)\ln(1-p)$$

- The first derivative of LL with respect to p is $U(p) = \dfrac{\partial LL}{\partial p} = \dfrac{\sum_{i=1}^{n} Y_i}{p} - \dfrac{n - \sum_{i=1}^{n} Y_i}{1-p}$

  And is referred to as the **score function**.

- To calculate the MLE of p, we set the score function, $U(p)$ equal to 0 and solve for p. In this case, we get an MLE of p that is:

$$\frac{\sum_{i=1}^{n} Y_i}{p} - \frac{n - \sum_{i=1}^{n} Y_i}{1-p} = \frac{(1-p)\sum_{i=1}^{n} Y_i}{p(1-p)} - \frac{p\left(n - \sum_{i=1}^{n} Y_i\right)}{p(1-p)} = \frac{\sum_{i=1}^{n} Y_i}{p(1-p)} - \frac{np}{p(1-p)} = 0$$

$$\frac{\sum_{i=1}^{n} Y_i}{p(1-p)} - \frac{np}{p(1-p)} = 0 \;\Rightarrow\; \sum_{i=1}^{n} Y_i - np = 0 \;\Rightarrow\; np = \sum_{i=1}^{n} Y_i$$

$$\Rightarrow\; \hat{p} = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}$$

Minimum or Maximum? (Bernoulli Example)

$$\frac{\partial^2 LL}{\partial p^2} = \frac{\partial\left[\frac{\sum_{i=1}^{n} Y_i}{p} - \frac{n-\sum_{i=1}^{n} Y_i}{1-p}\right]}{\partial p} = \frac{-\sum_{i=1}^{n} Y_i}{p^2} - \frac{\left(n-\sum_{i=1}^{n} Y_i\right)}{(1-p)^2}$$

- Evaluating at the MLE

$$\left(\frac{\partial^2 LL}{\partial p^2}\right)_{p=\hat{p}} = \frac{-\sum_{i=1}^{n} Y_i}{\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)^2} - \frac{\left(n-\sum_{i=1}^{n} Y_i\right)}{\left(1-\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)\right)^2} = \frac{-n^2}{\sum_{i=1}^{n} Y_i} - \frac{n^2}{\left(n-\sum_{i=1}^{n} Y_i\right)}$$

- When $0 < \sum_{i=1}^{n} Y_i < n$, the 2nd derivative at the MLE is negative
  - So the MLE is the maximum

- When $\sum_{i=1}^{n} Y_i = 0 \; or \; \sum_{i=1}^{n} Y_i = n \Rightarrow \hat{p} = 0 \; or \; \hat{p} = 1$ is said to be on the “boundary”

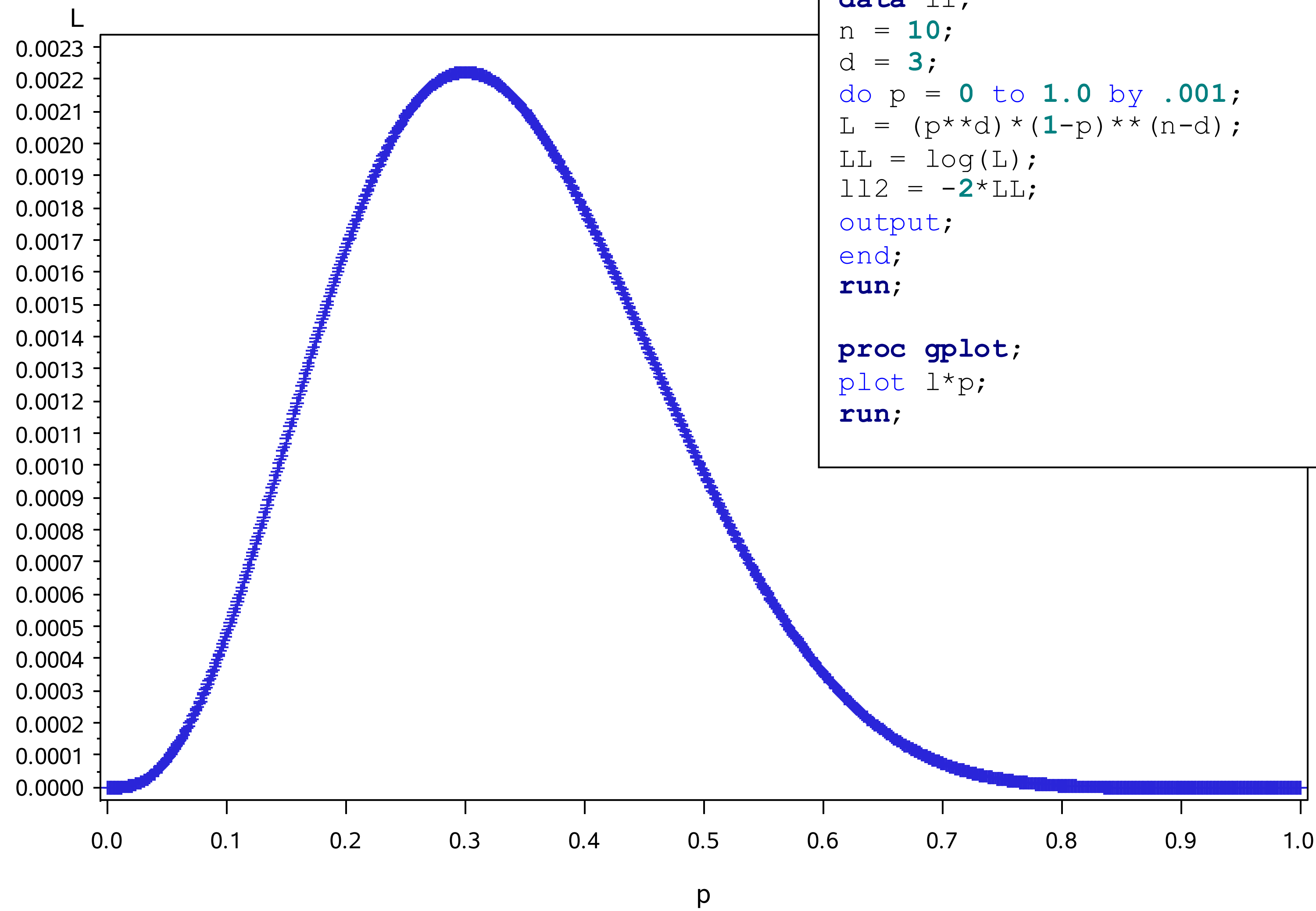
Maximum Likelihood Estimation (Bernoulli Example)

Example: Among 10 randomly selected individuals, 3 have a disease.

What is the MLE for *p*, the proportion in the population with the disease?

$$L = \prod_{i=1}^{n} p^{Yi}(1-p)^{1-Yi} \qquad \text{LL} = \left(\sum_{i=1}^{n} Y_i\right)\ln(p) + \left(n - \sum_{i=1}^{n} Y_i\right)\ln(1-p)$$

p	L	ln(L) = LL	-2LL
0	$0^3 1^7 = 0.0$	-∞	-∞
0.1	$0.1^3 0.9^7 = 0.00047$	-7.64527	15.2906
0.2	$0.2^3 0.8^7 = 0.001677$	-6.39032	12.7806
0.29	$0.29^3 0.71^7 = 0.002218$	-6.11106	12.2221
0.3=(# diseased)/n	$0.3^3 0.7^7 =$ **0.0022236**	**-6.10864**	**12.2173**
0.33	0.0022183	-6.12933	12.2587
0.4	0.00179159	-6.32465	12.6493
…			
1.0	0.0	-∞	-∞




```
data ll;
n = 10;
d = 3;
do p = 0 to 1.0 by .001;
L = (p**d)*(1-p)**(n-d);
LL = log(L);
ll2 = -2*LL;
output;
end;
run;

proc gplot;
plot l*p;
run;
```

Asymptotic Properties of MLEs

- The exact distribution of MLEs can be very complicated
  - Often have to rely on large sample methods instead

- Using a Taylor series expansion and the Delta Method, the following properties can be shown as $n \to \infty$

  1. $\hat{\theta}$ is asymptotically unbiased (i.e. $E(\hat{\theta}) \to \theta$)

     -However $\hat{\theta}$ may be biased for small or finite samples

  2. $\hat{\theta}$ is consistent

     -i.e. $\Pr\left\{\left|\hat{\theta}-\theta\right| > \varepsilon\right\} \to 0$

  3. $\hat{\theta}$ is asymptotically efficient

     -It achieves the minimum variance among all asymptotically unbiased estimators

- Using the central limit theorem, $\hat{\theta}$ is asymptotically normally distributed

  -i.e. $\hat{\theta} \sim N\left[\theta, Var\left(\hat{\theta}\right)\right]$

- How to calculate $\hat{Var}\left(\hat{\theta}\right)$?

Observed vs Expected Information

- How to calculate $V\hat{a}r\left(\hat{\theta}\right)$?

- Use the information function
  - The negative of the curvature in $LL = logL$.

1. The inverse of the expected information (Fisher Information)

$$V\hat{a}r\left(\hat{\theta}\right) = -\left\{E\left(\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right)\right\}^{-1}_{\theta=\hat{\theta}}$$

2. The inverse of the observed information

$$V\hat{a}r\left(\hat{\theta}\right) = -\left\{\left(\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right)\right\}^{-1}_{\theta=\hat{\theta}}$$

- In either case estimates of variance are obtained by evaluating the variance at the MLE

Observed vs Expected Information

- Estimates of variance calculated using either observed or expected information are similar for sufficiently large sample sizes
  - Cox DR and Snell EJ (1989) Analysis of Binary Data

- For the models we have considered (i.e. linear and logistic), the variance estimates constructed using either observed or expected information are identical

  - This equivalence occurs because all of these models are special cases of a broader family of generalized linear models
    - McCullagh P and Medler JA. (1989) Generalized Linear Models

- Efron and Hinkley (Biometrika, 1978; 65:457-487) have argued that, in general, better estimates of variance are obtained using observed rather than expected information

- As information increases (i.e. smaller variances) the log-likelihood becomes more peaked

Fisher Information (Bernoulli Example)

- Using the above MLE theory, for large n

$$\hat{\mathrm{p}} \sim N\left[\mathrm{p}, Var\left(\hat{p}\right)\right]$$

- For the likelihood considered previously, the expected information is:

$$I(\mathrm{p}) = E\left[-\left(\frac{\partial^2 LL}{\partial p^2}\right)\right] = E\left[-\frac{\partial}{\partial p}\left(\frac{\sum_{i=1}^{n} Y_i}{p} - \frac{n-\sum_{i=1}^{n} Y_i}{(1-p)}\right)\right] = E\left[\frac{\sum_{i=1}^{n} Y_i}{p^2} + \frac{n-\sum_{i=1}^{n} Y_i}{(1-p)^2}\right] = \left[\frac{\sum_{i=1}^{n} E[Y_i]}{p^2} + \frac{n-\sum_{i=1}^{n} E[Y_i]}{(1-p)^2}\right]$$

$$= \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p} + \frac{n}{(1-p)} = \frac{n(1-p)}{p(1-p)} + \frac{np}{p(1-p)} = \frac{n}{p(1-p)}$$

- To get the asymptotic variance, take the inverse

$$Var(\hat{\mathrm{p}}) = \left(\frac{n}{p(1-p)}\right)^{-1} = \frac{p(1-\mathrm{p})}{n} \Rightarrow \hat{p} \sim N\left(p, \frac{p(1-\mathrm{p})}{n}\right)$$

Bernoulli Example: Observed vs Expected Information

- The inverse of the expected information (Fisher Information)

$$V\hat{a}r\left(\hat{p}\right) = -\left\{E\left(\frac{\partial^2 \log L(\mathrm{p})}{\partial p^2}\right)\right\}_{p=\hat{p}}^{-1} = \left\{\frac{n}{p(1-p)}\right\}_{p=\hat{p}}^{-1} = \frac{\hat{p}\left(1-\hat{p}\right)}{n}$$

- The inverse of the observed information

$$V\hat{a}r\left(\hat{p}\right) = -\left\{\left(\frac{\partial^2 \log L(\mathrm{p})}{\partial p^2}\right)\right\}_{p=\hat{p}}^{-1} = \left\{\frac{\sum_{i=1}^{n} Y_i}{p^2} + \frac{\left(n - \sum_{i=1}^{n} Y_i\right)}{\left(1-p\right)^2}\right\}_{p=\hat{p}}^{-1} = \left\{\frac{n\hat{p}}{\hat{p}^2} + \frac{\left(n - n\hat{p}\right)}{\left(1-\hat{p}\right)^2}\right\}^{-1} = \left\{\frac{n\hat{p}}{\hat{p}^2} + \frac{n\left(1-\hat{p}\right)}{\left(1-\hat{p}\right)^2}\right\}^{-1}$$

$$= \left\{\frac{n}{\hat{p}} + \frac{n}{\left(1-\hat{p}\right)}\right\}^{-1} = \left\{\frac{n\left(1-\hat{p}\right)}{\hat{p}\left(1-\hat{p}\right)} + \frac{n\hat{p}}{\hat{p}\left(1-\hat{p}\right)}\right\}^{-1} = \left\{\frac{n}{\hat{p}\left(1-\hat{p}\right)}\right\}^{-1} = \frac{\hat{p}\left(1-\hat{p}\right)}{n}$$

- The observed and expected information are identical for **this example**

MLEs for coefficients in logistic regression

Recall that the logistic model takes the form:

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$

and solving for *p*:

$$p_i = \frac{e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}}} = \frac{e^{z_i}}{1+e^{z_i}} \quad \text{where} \quad z_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$

The likelihood for the *n* observations is:

$$L = \prod_{i=1}^{n} \overset{\overset{p_i}{\downarrow}}{\left(\frac{e^{z_i}}{1+e^{z_i}}\right)^{Y_i}} \overset{\overset{1-p_i}{\downarrow}}{\left(1 - \frac{e^{z_i}}{1+e^{z_i}}\right)^{1-Y_i}}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}}}\right)^{Y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}}}\right)^{1-Y_i}$$

$$= \prod_{i=1}^{n} \left(e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}}\right)^{Yi} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}}}\right)$$

MLEs for coefficients in logistic regression

- The p+1 score functions of β for the logistic regression model cannot by solved analytically.
    - It is common to use a numerical algorithm, such as the Newton-Raphson algorithm, to obtain the MLEs.

- The information in this case will be a (p+1) × (p+1) matrix of the partial second derivatives of l with respect to the parameters, β.
    - The inverted information matrix is the covariance matrix for $\hat{\beta}$.

And the log-likelihood is:

$$l = \sum_{i=1}^{n}\left[Y_i \log\left(\frac{e^{z_i}}{1+e^{z_i}}\right) + \left(1-Y_i\right)\log\left(1-\frac{e^{z_i}}{1+e^{z_i}}\right)\right]$$

Newton-Raphson Iteration

- Newton-Raphson iteration is a technique used to find roots of equations
  - e.g., any real number r is called a root for the equation f(r) = 0

- Use Newton-Raphson iteration to maximize the log-likelihood by obtaining estimators of regression coefficients (i.e., MLE's) which are roots of the score functions

- Application of Newton-Raphson to obtain MLE's requires inverting the matrix of second derivatives of the log-likelihood

- Alternative numerical methods are available which neither calculate the matrix of second derivatives of the log-likelihood nor its inverse

How does Newton-Raphson iteration find the root of f(r) = 0?

1. Select a starting value for r, $X_{(0)}$

2. Approximate the point f (r) using a first order Taylor series expansion

$$f(r) \approx f(X_{(0)}) + f'(X_{(0)})(X - X_{(0)})$$

3. Set the linear approximation equal to zero and solve for X

$$X_{(1)}=X_{(0)} - [f(X_{(0)})]/[f'(X_{(0)})]$$

4. Iterate until convergence

- Newton-Raphson iteration will converge very quickly if
  - f, f′, and f″ are continuous in a neighborhood of a root r of f
  - f′(r) does not equal 0
  - And X(0) has been well chosen

- For more on this topic, please see the following blog

http://thelaziestprogrammer.com/sharrington/math-of-machine-learning/solving-logreg-newtons-method