e. Example: $X_1, \cdots, X_n \sim \mathcal{N}(\mu, 1)$.

i. $\tilde{\theta} = X_1$.

ii. $\bar{X}$ sufficient

iii. $\text{Cov}\left[X_1, \bar{X}\right] = 1/n$

iv. $(\bar{X}, X_1) \sim \mathcal{N}((\mu,\mu), \begin{pmatrix} 1/n & 1/n \\ 1/n & 1 \end{pmatrix})$.

v. Inverse of variance matrix is $\begin{pmatrix} \frac{n^2}{n-1} & -\frac{n}{n-1} \\ -\frac{n}{n-1} & \frac{n}{n-1} \end{pmatrix}$

vi. So joint density of $(U = \bar{X}, V = X_1)$ is

$$\frac{\exp(-\frac{n((u-\mu)^2 n - 2(u-\mu)(v-\mu) + (v-\mu)^2)}{2(n-1)}))n}{2\pi\sqrt{n-1}}$$

$$= n\exp(-(n(u-\mu)^2 + (v-u)^2(n/(n-1)))/2)/(2\pi\sqrt{n-1}).$$

vii. Then $\mathrm{E}\left[X_1|\bar{X}\right] = \bar{X}$.

f. Repeating Rao-Blackwellizing a second time doesn't change the estimator,

i. because of $\tilde{\theta}$ is a function only of the sufficient statistic U , then $\mathrm{E}\left[\tilde{\theta}(U)|U\right] = \tilde{\theta}(U)$.

ii. If sufficient statistic U is of the same dimensionality of θ , then with some additional conditions can show that Rao-Blackwellizing an unbiased estimator gives you the smallest possible variance.

iii. Hence Rao-Blackwellizing gives *Minimum Variance Unbiased Estimator* (MVUE).

WMS: 9,8

P. Properties of the MLE:

1. Score function has expectation zero

a. Score function is $U = \ell'(\theta; \boldsymbol{X}) = \frac{d}{d\theta} f_{\boldsymbol{X}}(\boldsymbol{X}, \theta)$

b. $1 = \int \exp(\ell(\theta; \boldsymbol{x}))\ d\boldsymbol{x}$ (note integral is over all of data space).

c. $0 = \int \frac{d}{d\theta}\exp(\ell(\theta; \boldsymbol{x}))\ d\boldsymbol{x} = t\ell'(\theta)\exp(\ell(\theta; \boldsymbol{x}))\ d\boldsymbol{x} = \mathrm{E}\left[U\right]$

2. If data are IID, score function is a sample sum

a. $\ell(\theta) = \log(\prod_{i=1}^n f_{X_i}(X_i;\theta)) = \sum_{i=1}^n \log(f_{X_i}(X_i;\theta))$

b. $\ell'(\theta) = \sum_{i=1}^n U_i$ for $U_i = \frac{d}{d\theta} f_{X_i}(X_i;\theta)/f_{X_i}(X_i;\theta)$

c. Also, $\ell''(\theta) = \sum_{i=1}^n \frac{d^2}{d\theta^2}\log(f_{X_i}(X_i;\theta)) \approx -n\mathrm{E}\left[-\frac{d^2}{d\theta^2}\log(f_{X_i}(X_i;\theta))\right]$.

3. If $i(\theta) = \mathrm{E}\left[-\frac{d^2}{d^2\theta}\ln f_{X_i}(X_i;\theta)\right] < \infty$, then $i(\theta) = \mathrm{Var}\left[U_i\right]$

a. Proved as part of proof of Cramer-Rao lower bound.

4. $\hat{\theta} \approx \theta - [\ell''(\theta)]^{-1}U$.

a. Expanding $\ell'(\theta)$ as a Taylor series, $0 = \ell'(\hat{\theta}) \approx \ell'(\theta) + \ell''(\theta)(\hat{\theta} - \theta)$.

b. Solve for θ .

5. Hence $\hat{\theta}$ is approximately unbiased, with variance approximately $i(\theta)^{-1}/n$.

6. See Fig. 7.

a. Approximately hits the CR lower bound.

7. Same calculations hold for vector parameters.

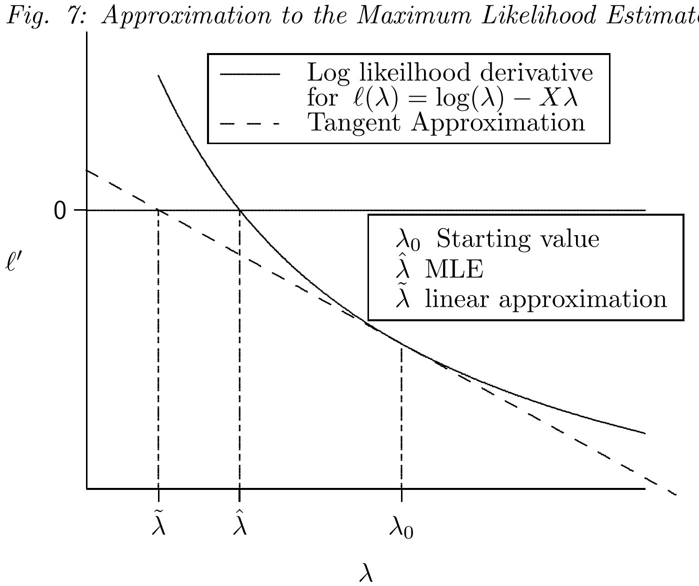
a. CLT applies to U .

8. Example: $X_1, \cdots, X_n\ N(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma)$.

a. $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n[-\theta_2^{-2}(X_i - \theta_1)^2/2 - \log(2\pi) - \log(\theta_2)]$

b. $\ell'(\boldsymbol{\theta}) = \sum_{i=1}^n((X_i - \theta_1)\theta_2^{-2}, \theta_2^{-3}(X_i - \theta_1)^2 - 1/\theta_2)$

Fig. 7: Approximation to the Maximum Likelihood Estimate



Exponential Dist., $\lambda_0 = 1$. Data $X = 1.5$. MLE $\hat{\lambda} = 2/3$

i. $U_i = ((X_i - \theta_1)\theta_2^{-2}, \theta_2^{-3}(X_i - \theta_1)^2 - 1/\theta_2)$.

c. $\ell'(\hat{\boldsymbol{\theta}}) = \mathbf{o}$ implies $\hat{\theta}_1 = \bar{X}$, $\hat{\theta}_2 = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/n}$, as before.

d. $\ell''(\boldsymbol{\theta}) = \sum_{i=1}^n \begin{pmatrix} -\theta_2^{-2} & -2(X_i - \theta_1)\theta_2^{-3} \\ -2(X_i - \theta_1)\theta_2^{-3} & -3\theta_2^{-4}(X_i - \theta_1)^2 + \theta_2^{-2} \end{pmatrix}$.

e. $\ell''(\hat{\boldsymbol{\theta}}) = -\begin{pmatrix} n\hat{\theta}_2^{-2} & 0 \\ 0 & 3n\hat{\theta}_2^{-2} - n\hat{\theta}_2^{-2} \end{pmatrix} = -\begin{pmatrix} n\hat{\theta}_2^{-2} & 0 \\ 0 & 2n\hat{\theta}_2^{-2} \end{pmatrix}$.

f. Information is

$$i(\theta) = \mathrm{E}\left[-\begin{pmatrix} -\theta_2^{-2} & -2(X_i - \theta_1)\theta_2^{-3} \\ -2(X_i - \theta_1)\theta_2^{-3} & -3\theta_2^{-4}(X_i - \theta_1)^2 + \theta_2^{-2} \end{pmatrix}\right] = \begin{pmatrix} \theta_2^{-2} & 0 \\ 0 & 2\theta_2^{-2} \end{pmatrix}.$$

g. Hence $\mathrm{Var}\left[\hat{\boldsymbol{\theta}}\right] = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{pmatrix}$.

h. See Table 2 for values of variance.

Table 2: Asymptotic and True Variance for Sample Standard Deviation

n	Asymptotic Variance $1/(2n)$	Expect-ation	True Variance
2.	0.2500	0.56419	0.18169
3.	0.1667	0.72360	0.14307
4.	0.1250	0.79789	0.11338
5.	0.1000	0.84075	0.09314
6.	0.0833	0.86862	0.07882
7.	0.0714	0.88820	0.06824
8.	0.0625	0.90270	0.06013
9.	0.0556	0.91388	0.05372
10.	0.0500	0.92275	0.04854

9. Example: $Y_i \sim \mathcal{B}\text{in}(\pi_i, 1)$ for $\pi_i = \exp(\beta_1 + \beta_2 x_i)/(1 + \exp(\beta_1 + \beta_2 x_i))$, independent.

a. $$\mathrm{P}\left[Y_i = y_i\right] = \prod_{i=1}^{n} \frac{(\exp(\beta_1 + \beta_2 x_i))^{y_i}}{(1 + \exp(\beta_1 + \beta_2 x_i))}$$
$$= \frac{\exp(\beta_1 \sum_{i=1}^{n} y_i + \beta_2 \sum_{i=1}^{n} x_i y_i)}{\prod_{i=1}^{n}(1 + \exp(\beta_1 + \beta_2 x_i))}$$

i. Hence $(U, V) = (\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i y_i)$ is sufficient

b. $\ell(\beta_1, \beta_2) = \beta_1 U + \beta_2 V - \sum_{i=1}^{n} \log(1 + \exp(\beta_1 + \beta_2 x_i))$

c. $\ell'(\beta_1, \beta_2) = (U, V) - \sum_{i=1}^{n} \frac{\exp(\beta_1+\beta_2 x_i)(1, x_i)}{(1+\exp(\beta_1+\beta_2 x_i))} = (U, V) - \sum_{i=1}^{n} \pi_i (1, x_i)$

d. $\ell''(\beta_1, \beta_2) = -\sum_{i=1}^{n} \pi_i'(1, x_i)$

i. π_i' is
$$\frac{(1+\exp(\beta_1+\beta_2 x_i))\exp(\beta_1+\beta_2 x_1) - \exp(\beta_1+\beta_2 x_i)\exp(\beta_1+\beta_2 x_i)}{(1+\exp(\beta_1+\beta_2 x_i))^2}(1, x_i)$$
$= \pi_i(1 - \pi_i)(1, x_i)$

ii. $\ell''(\beta_1, \beta_2) = -\sum_{i=1}^{n} \pi_i(1 - \pi_i)(1, x_i)(1, x_i)^\top$

iii. $\hat{\boldsymbol{\beta}}$ solves $\sum_{i=1}^{n} \hat{\pi}_i(1, x_i) = (U, V)$.

iv. Distribution is approximately $\mathcal{N}(\boldsymbol{\beta}, [\sum_{i=1}^{n} \pi_i(1 - \pi_i)(1, x_i)(1, x_i)^\top]^{-1})$

10. Interesting aspects of this example:

a. $\hat{\boldsymbol{\beta}}$ generally not available in closed form

i. Algorithm for fitting $\hat{\boldsymbol{\beta}}$ uses the Talor series areguemnt as before, with previous updates for $\hat{\boldsymbol{\beta}}$ in place of true value $\boldsymbol{\beta}$.

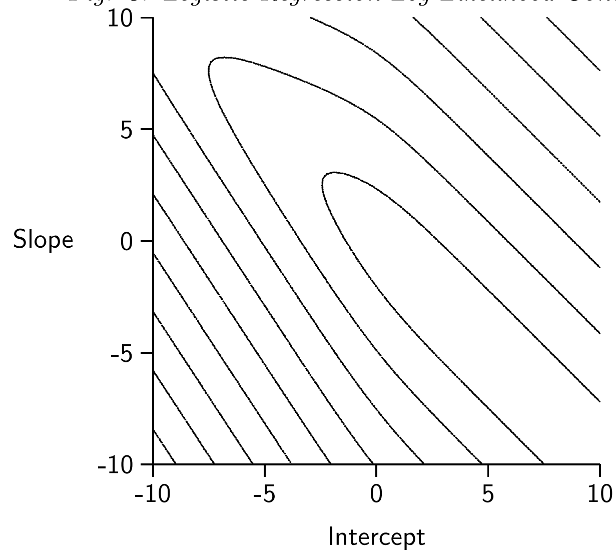
ii. Algorithm is called Newton-Raphson generally

iii. Fisher Scoring in statistics lingo.

iv. Only guaranteed to work when the log likelihood is concave, except

v. See Fig. 8.

Fig. 8: Logistic Regression Log Likelihood Contours



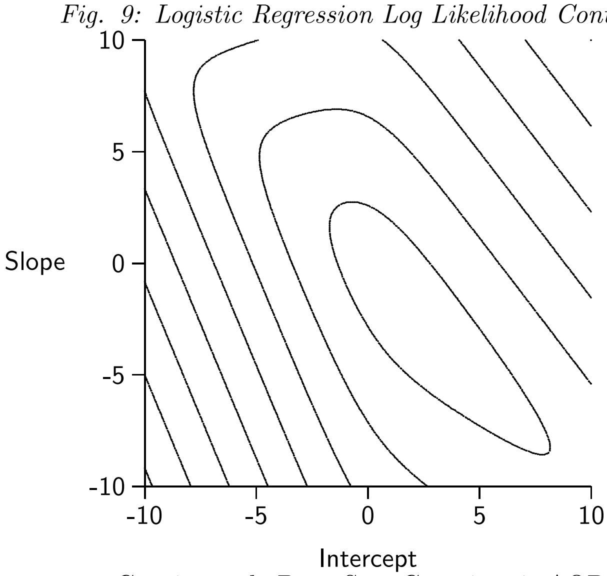
Goorin et al. Data Set. Covariate is LI

b. For some collections of responses Y_i , mle is not finite.

i. See Fig. 9.

05

Fig. 9: Logistic Regression Log Likelihood Contours



Goorin et al. Data Set. Covariate is AOP

This page intentionally left blank.