STA 6351, Report.2.9

Carson Slater *Baylor University*

2.9

Suppose $Y_i \sim$ Bernoulli(p_i) with logit$(p_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, $i = 1, \ldots, n$, where $\mathbf{X}$ is the $n \times p$ design matrix.

(a) Write down the log-likelihood function $l(\boldsymbol{\beta}; \mathbf{y})$.

The likelihood function is the product of the individual Bernoulli probability mass functions:

$$L(\boldsymbol{\beta}; \mathbf{y}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}$$

The log-likelihood function $l(\boldsymbol{\beta}; \mathbf{y})$ is the natural logarithm of the likelihood:

$$l(\boldsymbol{\beta}; \mathbf{y}) = \log L(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

We can re-express this by factoring out y_i:

$$\begin{aligned} l(\boldsymbol{\beta}; \mathbf{y}) &= \sum_{i=1}^{n} [y_i \log(p_i) - y_i \log(1 - p_i) + \log(1 - p_i)] \\ &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{p_i}{1 - p_i} \right) + \log(1 - p_i) \right] \end{aligned}$$

Given the logit link, $\text{logit}(p_i) = \log(p_i/(1 - p_i)) = \mathbf{x}_i^\top \boldsymbol{\beta}$. Let $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$. From $\eta_i = \log(p_i/(1 - p_i))$, we exponentiate to get $e^{\eta_i} = p_i/(1 - p_i)$, which implies $p_i = e^{\eta_i}(1 - p_i) \Rightarrow p_i(1 + e^{\eta_i}) = e^{\eta_i} \Rightarrow p_i = e^{\eta_i}/(1 + e^{\eta_i})$. This also gives $1 - p_i = 1 - \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{\eta_i}}$. Taking the log, we find $\log(1 - p_i) = \log(1/(1 + e^{\eta_i})) = -\log(1 + e^{\eta_i})$. Substituting $\log(p_i/(1 - p_i)) = \eta_i$ and $\log(1 - p_i) = -\log(1 + e^{\eta_i})$ into the log-likelihood equation:

$$\begin{aligned} l(\boldsymbol{\beta}; \mathbf{y}) &= \sum_{i=1}^{n} [y_i \eta_i - \log(1 + e^{\eta_i})] \\ &= \sum_{i=1}^{n} \left[y_i (\mathbf{x}_i^\top \boldsymbol{\beta}) - \log(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}) \right] \end{aligned}$$

(b) Differentiate to obtain the score vector $U(\boldsymbol{\beta}) = \mathbf{X}^\top(\mathbf{y} - \mathbf{p}(\boldsymbol{\beta}))$ and $H(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{X}$, respectively, where $\mathbf{W}(\boldsymbol{\beta}) =$ diag$(p_i(\boldsymbol{\beta})(1 - p_i(\boldsymbol{\beta})))$.

The score vector $U(\boldsymbol{\beta})$ is the $p \times 1$ vector of first partial derivatives of $l(\boldsymbol{\beta}; \mathbf{y})$ with respect to $\boldsymbol{\beta}$. We use the chain rule, differentiating with respect to $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$:

$$U(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\frac{\partial l_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \right)$$

First, $\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \frac{\partial (\mathbf{x}_i^\top \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{x}_i$. Second, $\frac{\partial l_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i} \left(y_i \eta_i - \log(1 + e^{\eta_i}) \right) = y_i - \frac{1}{1+e^{\eta_i}} \cdot e^{\eta_i} = y_i - \frac{e^{\eta_i}}{1+e^{\eta_i}}$. Since $p_i = e^{\eta_i}/(1 + e^{\eta_i})$, we have $\frac{\partial l_i}{\partial \eta_i} = y_i - p_i$. Substituting back, $U(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - p_i)\mathbf{x}_i$. In matrix form, this is $U(\boldsymbol{\beta}) = \mathbf{X}^\top (\mathbf{y} - \mathbf{p}(\boldsymbol{\beta}))$.

The Hessian matrix $H(\boldsymbol{\beta})$ is the $p \times p$ matrix of second partial derivatives, found by differentiating $U(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}^\top$:

$$\begin{aligned} H(\boldsymbol{\beta}) &= \frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \left(\sum_{i=1}^{n} (y_i - p_i)\mathbf{x}_i \right) \\ &= \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial (y_i - p_i)}{\partial \boldsymbol{\beta}^\top} = \sum_{i=1}^{n} \mathbf{x}_i \left(-\frac{\partial p_i}{\partial \boldsymbol{\beta}^\top} \right) \end{aligned}$$

We apply the chain rule to $\frac{\partial p_i}{\partial \boldsymbol{\beta}^\top}$:

$$\frac{\partial p_i}{\partial \boldsymbol{\beta}^\top} = \frac{\partial p_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^\top} = \frac{\partial p_i}{\partial \eta_i} \mathbf{x}_i^\top$$

We need the derivative of p_i with respect to η_i. Using the quotient rule:

$$\begin{aligned} \frac{\partial p_i}{\partial \eta_i} &= \frac{\partial}{\partial \eta_i} \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right) = \frac{e^{\eta_i}(1 + e^{\eta_i}) - e^{\eta_i}(e^{\eta_i})}{(1 + e^{\eta_i})^2} \\ &= \frac{e^{\eta_i} + e^{2\eta_i} - e^{2\eta_i}}{(1 + e^{\eta_i})^2} = \frac{e^{\eta_i}}{(1 + e^{\eta_i})^2} \\ &= \left(\frac{e^{\eta_i}}{1 + e^{\eta_i}} \right) \left(\frac{1}{1 + e^{\eta_i}} \right) = p_i(1 - p_i) \end{aligned}$$

This is the variance of a Bernoulli(p_i) variable. Let $w_i = p_i(1 - p_i)$. Substituting back: $\frac{\partial p_i}{\partial \boldsymbol{\beta}^\top} = w_i \mathbf{x}_i^\top$. Finally, substituting this into the expression for the Hessian:

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i (-w_i \mathbf{x}_i^\top) = -\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^\top$$

This sum of outer products can be written in matrix form. Let $\mathbf{W}(\boldsymbol{\beta})$ be the $n \times n$ diagonal matrix with $w_i = p_i(\boldsymbol{\beta})(1 - p_i(\boldsymbol{\beta}))$ on the diagonal. Then:

$$H(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}$$

(c) For a reference point $\boldsymbol{\beta}_0$, show that a first-order Taylor expansion yields

$$\frac{L(\boldsymbol{\beta}_1)}{L(\boldsymbol{\beta}_0)} \approx \exp\{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0)\}.$$

Which aggregated sample sums appear inside $U(\boldsymbol{\beta}_0)$?

We perform a first-order Taylor expansion of the log-likelihood function $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$ around the point $\boldsymbol{\beta}_0$:

$$l(\boldsymbol{\beta}_1) \approx l(\boldsymbol{\beta}_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top \nabla l(\boldsymbol{\beta}_0)$$

The gradient $\nabla l(\boldsymbol{\beta}_0)$ is by definition the score vector $U(\boldsymbol{\beta}_0)$.

$$l(\boldsymbol{\beta}_1) \approx l(\boldsymbol{\beta}_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0)$$

Rearranging the terms gives the log-likelihood ratio:

$$l(\boldsymbol{\beta}_1) - l(\boldsymbol{\beta}_0) \approx (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0)$$

Since $l(\boldsymbol{\beta}_1) - l(\boldsymbol{\beta}_0) = \log L(\boldsymbol{\beta}_1) - \log L(\boldsymbol{\beta}_0) = \log(L(\boldsymbol{\beta}_1)/L(\boldsymbol{\beta}_0))$:

$$\log \frac{L(\boldsymbol{\beta}_1)}{L(\boldsymbol{\beta}_0)} \approx (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0)$$

Exponentiating both sides gives the desired approximation for the likelihood ratio:

$$\frac{L(\boldsymbol{\beta}_1)}{L(\boldsymbol{\beta}_0)} \approx \exp\{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0)\}$$

The score vector $U(\boldsymbol{\beta}_0)$ is $U(\boldsymbol{\beta}_0) = \mathbf{X}^\top(\mathbf{y} - \mathbf{p}(\boldsymbol{\beta}_0)) = \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{p}_0$. The aggregated sample sums that appear inside $U(\boldsymbol{\beta}_0)$ are the elements of the $p \times 1$ vector $\mathbf{X}^\top\mathbf{y}$. The j-th element of this vector is $\sum_{i=1}^n x_{ij}y_i$, which is the sum of the j-th predictor variable weighted by the observed outcomes y_i.

(d) Extend to second order, verifying that

$$\log \frac{L(\boldsymbol{\beta}_1)}{L(\boldsymbol{\beta}_0)} \approx (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0) - \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top \mathbf{X}^\top \mathbf{W}_0 \mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0),$$

where $\mathbf{W}_0 = \mathbf{W}(\boldsymbol{\beta}_0)$.

We extend the Taylor expansion of $l(\boldsymbol{\beta}_1)$ around $\boldsymbol{\beta}_0$ to the second order. A second-order expansion of a vector function $f(\mathbf{z})$ around $\mathbf{z}_0$ is $f(\mathbf{z}_1) \approx f(\mathbf{z}_0) + (\mathbf{z}_1 - \mathbf{z}_0)^\top \nabla f(\mathbf{z}_0) + \frac{1}{2}(\mathbf{z}_1 - \mathbf{z}_0)^\top H_f(\mathbf{z}_0)(\mathbf{z}_1 - \mathbf{z}_0)$, where H_f is the Hessian matrix. Applying this to $l(\boldsymbol{\beta})$:

$$l(\boldsymbol{\beta}_1) \approx l(\boldsymbol{\beta}_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top \nabla l(\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top H(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

From part (b), we know the gradient is the score vector $\nabla l(\boldsymbol{\beta}_0) = U(\boldsymbol{\beta}_0)$ and the Hessian is $H(\boldsymbol{\beta}_0) = -\mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta}_0)\mathbf{X} = -\mathbf{X}^\top \mathbf{W}_0 \mathbf{X}$. Substituting these into the expansion:

$$l(\boldsymbol{\beta}_1) \approx l(\boldsymbol{\beta}_0) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top (-\mathbf{X}^\top \mathbf{W}_0 \mathbf{X})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

Subtracting $l(\boldsymbol{\beta}_0)$ from both sides gives the quadratic approximation to the log-likelihood ratio:

$$l(\boldsymbol{\beta}_1) - l(\boldsymbol{\beta}_0) \approx (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0) - \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top \mathbf{X}^\top \mathbf{W}_0 \mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

Recognizing $l(\boldsymbol{\beta}_1) - l(\boldsymbol{\beta}_0) = \log(L(\boldsymbol{\beta}_1)/L(\boldsymbol{\beta}_0))$, we have verified the relationship.

(e) How do the score vector and observed information act as an approximate sufficient summary of the data near $\boldsymbol{\beta}_0$?

The result from (d) shows that the log-likelihood ratio $\log(L(\boldsymbol{\beta}_1)/L(\boldsymbol{\beta}_0))$ for $\boldsymbol{\beta}_1$ near $\boldsymbol{\beta}_0$ is approximately:

$$\log \frac{L(\boldsymbol{\beta}_1)}{L(\boldsymbol{\beta}_0)} \approx (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top U(\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^\top H(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

This is a quadratic function of $(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$. This approximation of the log-likelihood surface depends on the data $\mathbf{y}$ only through the score vector $U(\boldsymbol{\beta}_0) = \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{p}_0$. The other term, the Hessian $H(\boldsymbol{\beta}_0) = -\mathbf{X}^\top \mathbf{W}_0 \mathbf{X}$, depends only on the reference point $\boldsymbol{\beta}_0$ and the design matrix $\mathbf{X}$, not on the n-dimensional data vector $\mathbf{y}$. The observed information matrix is $I(\boldsymbol{\beta}_0) = -H(\boldsymbol{\beta}_0) = \mathbf{X}^\top \mathbf{W}_0 \mathbf{X}$. Therefore, for local inference about $\boldsymbol{\beta}$ near $\boldsymbol{\beta}_0$, the entire n-dimensional data vector $\mathbf{y}$ influences the shape of the likelihood surface only through the p-dimensional score vector $U(\boldsymbol{\beta}_0)$. By the logic of the Factorization Theorem, this implies that the score vector $U(\boldsymbol{\beta}_0)$, evaluated at $\boldsymbol{\beta}_0$, along with the (data-independent) observed information $I(\boldsymbol{\beta}_0)$, contains all the relevant information from the sample for making inferences about $\boldsymbol{\beta}$ in this local neighborhood. They act as an *approximate sufficient summary* (or "approximate sufficient statistic") of the full data $\mathbf{y}$, reducing the n-dimensional data to a p-dimensional summary for the purpose of local inference.