Lecture 2: Maximum Likelihood and Friends

Chris Conlon

February 6, 2021

NYU Stern

Examples

We are going to practice writing down the:

- Likelihood $L(\theta) = \Pr(z_1, \ldots, z_n; \theta) = \prod_{i=1}^{N} f(z_i; \theta)$.
- log likelihood $\ell(\theta) = \sum_{i=1}^{N} \ln f(z_i; \theta) = \sum_{i=1}^{N} \ell_i(z_i; \theta)$.
- Scores $\mathcal{S}_i(z_i, \theta) = -\frac{\partial \ln f(z_i; \theta)}{\partial \theta} = \frac{\partial \ell_i(z_i; \theta)}{\partial \theta}$.
- Hessian contribution $\mathcal{H}_i(z_i, \theta) = \frac{\partial^2 \ell_i(z_i; \theta)}{\partial \theta \partial \theta'}$.
- Information Matrix $\mathcal{I}(z_i, \theta) = \mathbb{E}_z[-\mathcal{H}_i(z_i, \theta)] = \mathbb{E}_z[\mathcal{S}_i(z_i, \theta)\mathcal{S}_i(z_i, \theta)^T]$
- Variance $V(\theta) \geq [\mathcal{I}(z_i, \theta)]^{-1}$ (Cramer-Rao Lower Bound).

Exponential Example

- Suppose we have data that are expoentially distributed $Y_i \sim \text{Exp}(\lambda_i)$. The goal is to estimate the parameter λ_i via MLE and $f(y_i|\lambda_i) = \lambda e^{-\lambda y_i}$.
- Sometimes we want to parameterize the rate $\lambda_i = x_i'\beta$ with covariates.
- Example: Time until next customer arrives varies with time of day, day of week, etc. Time until default might depend on credit score, debt-to-income, market conditions, etc.

Exponential Regression

start with pdf:

$$f_{Y|X}(y|x,\beta) = x'\beta \exp\left(-y \cdot x'\beta\right)$$

then log likelihood

$$\ell(\beta) = \sum_{i=1}^{N} \ln f_{Y|X}\left(y_i|x_i,\beta\right) = \sum_{i=1}^{N} \ln(x_i'\beta) - y_i \cdot \left(x_i'\beta\right)$$

And Score, Hessian:

$$\mathcal{S}_i(y_i, x_i, \beta) = x_i'\left(\frac{1}{x_i'\beta} - y_i\right)$$

$$\mathcal{H}(y, x, \beta) = \mathbb{E}\left[x_i' x_i \left(\frac{1}{x_i'\beta}\right)^2\right]$$

Logit/Probit Example

- Can construct an MLE:

$$\begin{aligned}\hat{\beta}^{MLE} &= \arg\max_{\beta} \prod_{i=1}^{N} F(Z_i)^{y_i}(1 - F(Z_i))^{1-y_i} \\ Z_i &= \beta_0 + \beta_1 X_i\end{aligned}$$

- Probit: $F(Z_i) = \Phi(Z_i)$ and its derivative (density) $f(Z_i) = \phi(Z_i)$. Also is symmetric so that $1 - F(Z_i) = F(-Z_i)$.
- Logit: $F(Z_i) = \frac{1}{1+e^{-z}}$ and its derivative (density) $f(Z_i) = \frac{e^{-z}}{(1+e^{-z})^2}$ a more convenient property is that $\frac{f(z)}{F(z)} = 1 - F(z)$ this is called the hazard rate.

A probit trick

Let $q_i = 2y_i - 1$

$$F(q_i \cdot Z_i) = \begin{cases} F(Z_i) & \text{when } y_i = 1 \\ F(-Z_i) = 1 - F(Z_i) & \text{when } y_i = 0 \end{cases}$$

So that

$$\ell(y_1, \ldots, y_n | \beta) = \sum_{i=1}^{N} \ln F(q_i \cdot Z_i)$$

FOC of Log-Likelihood

$$
\begin{aligned}
\ell(y_1, \ldots, y_n|\beta) &= \sum_{i=1}^{N} y_i \ln F(Z_i) + (1 - y_i) \ln(1 - F(Z_i)) \\
\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{N} \frac{y_i}{F(Z_i)} \frac{dF}{d\beta}(Z_i) - \frac{1 - y_i}{1 - F(Z_i)} \frac{dF}{d\beta}(Z_i) \\
&= \sum_{i=1}^{N} \frac{y_i \cdot f(Z_i)}{F(Z_i)} \frac{dZ_i}{d\beta} - \sum_{i=1}^{N} \frac{(1 - y_i) \cdot f(Z_i)}{1 - F(Z_i)} \frac{dZ_i}{d\beta} \\
&= \sum_{i=1}^{N} \left[\frac{y_i \cdot f(Z_i)}{F(Z_i)} X_i - \frac{(1 - y_i) \cdot f(Z_i)}{1 - F(Z_i)} X_i \right]
\end{aligned}
$$

FOC of Log-Likelihood (Logit)

This is the score of the log-likelihood:

$$\frac{\partial \ell}{\partial \beta} = \nabla_\beta \cdot \ell(\mathbf{y};\beta) \quad = \quad \sum_{i=1}^{N} \left[y_i \frac{f(Z_i)}{F(Z_i)} - (1 - y_i) \frac{f(Z_i)}{1 - F(Z_i)} \right] \cdot X_i$$

It is technically also a moment condition. It is easy for the logit

$$\begin{aligned} \nabla_\beta \cdot \ell(\mathbf{y};\beta) &= \sum_{i=1}^{N} [y_i(1 - F(Z_i)) - (1 - y_i) F(Z_i)] \cdot X_i \\ &= \sum_{i=1}^{N} \underbrace{[y_i - F(Z_i)]}_{\varepsilon_i} \cdot X_i \end{aligned}$$

This comes from the hazard rate.

FOC of Log-Likelihood (Probit)

This is the score of the log-likelihood:

$$
\begin{aligned}
\frac{\partial l}{\partial \beta} = \nabla_\beta \cdot \ell(\mathbf{y};\beta) &= \sum_{i=1}^{N}\left[y_i \frac{f(Z_i)}{F(Z_i)} - (1-y_i)\frac{f(Z_i)}{1-F(Z_i)}\right]\cdot X_i \\
&= \sum_{y_i=1} \frac{\phi(Z_i)}{\Phi(Z_i)} X_i + \sum_{y_i=0} \frac{-\phi(Z_i)}{1-\Phi(Z_i)} X_i
\end{aligned}
$$

Using the $q_i = 2y_i - 1$ trick

$$
\nabla_\beta \cdot \ell(\mathbf{y};\beta) = \sum_{i=1}^{N} \underbrace{\frac{q_i \phi(q_i Z_i)}{\Phi(Z_i)}}_{\lambda_i} X_i
$$

The Hessian Matrix

We could also take second derivatives to get the Hessian matrix:

$$\frac{\partial \ell^2}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} y_i \frac{f(Z_i)f(Z_i) - f'(Z_i)F(Z_i)}{F(Z_i)^2} X_i X_i'$$
$$+\sum_{i=1}^{N}(1 - y_i)\frac{f(Z_i)f(Z_i) - f'(Z_i)(1 - F(Z_i))}{(1 - F(Z_i))^2} X_i X_i'$$

This is a $K \times K$ matrix where K is the dimension of X or β.

The Hessian Matrix (Logit)

For the logit this is even easier (use the simplified logit score):

$$
\begin{aligned}
\frac{\partial \ell^2}{\partial \beta \partial \beta'} &= -\sum_{i=1}^{N} f(Z_i) X_i X_i' \\
&= -\sum_{i=1}^{N} F(Z_i)(1 - F(Z_i)) X_i X_i'
\end{aligned}
$$

This is negative semi definite

The Hessian Matrix (Probit)

Recall

$$\nabla_{\beta} \cdot \ell(\mathbf{y}; \beta) = \sum_{i=1}^{N} \underbrace{\frac{q_i \phi(q_i Z_i)}{\Phi(Z_i)}}_{\lambda_i} X_i$$

Take another derivative and recall $\phi'(z_i) = -z_i \phi(z_i)$

$$\begin{aligned} \nabla_{\beta}^2 \cdot \ell(\mathbf{y}; \beta) &= \sum_{i=1}^{N} \frac{q_i \phi'(q_i Z_i)\Phi(z_i) - q_i \phi(z_i)^2}{\Phi(z_i)^2} X_i X_i' \\ &= -\lambda_i (z_i + \lambda_i) \cdot X_i X_i' \end{aligned}$$

Hard to show but this is negative definite too.

Inference

- If we have the Hessian Matrix, inference is straightforward.
- $\mathcal{H}_f(\hat{\beta}^{MLE})$ tells us about the curvature of the log-likelihood around the maximum.
    - Function is flat $\rightarrow$ not very precise estimates of parameters
    - Function is steep $\rightarrow$ precise estimates of parameters
- Construct Fisher Information $\mathcal{I}(\hat{\beta}^{MLE}) = \mathbb{E}[-\mathcal{H}_f(\hat{\beta}^{MLE})]$ where expectation is over the data.
    - Logit does not depend on y_i so $\mathbb{E}[\mathcal{H}_f(\hat{\beta}^{MLE})] = \mathcal{H}_f(\hat{\beta}^{MLE})$.
    - Probit does depend on y_i so $\mathbb{E}[\mathcal{H}_f(\hat{\beta}^{MLE})] \neq \mathcal{H}_f(\hat{\beta}^{MLE})$.
- Inverse Fisher information $\mathbb{E}[-\mathcal{H}_f(\hat{\beta}^{MLE})]^{-1}$ is an estimate of the variance covariance matrix for $\hat{\beta}$.
- $\sqrt{diag[\mathbb{E}[-\mathcal{H}_f(\hat{\beta}^{MLE})]^{-1}]}$ is an estimate for $SE(\hat{\beta})$.

Thanks!