Fitting logistic regression models

Announcements

- Office hour times:
  - Monday 3 - 4 (sign up for 15-minute slots)
  - Wednesday 11 - 12 (15-minute slots)
  - Wednesday 12 - 12:45 (drop-in)
  - Thursday 1 - 2 (drop-in)
- Homework 1 and Challenge Assignment 1 released on course website

Course components

- Regular homework assignments
  - Practice material from class
- Challenge assignments
  - Learn additional material related to course
- 2 take-home exams
  - Demonstrate knowledge of theory and methodology
  - No final exam!
- 2 projects
  - Apply material to real data and real research questions

Assigning grades: specifications grading

To get a **B** in the course:

- Receive credit for at least 5 homework assignments
- Master one project
- Master at least 80% of the questions on both exams

To get an **A** in the course:

- Receive credit for at least 5 homework assignments
- Master both projects
- Master at least 80% of the questions on both exams
- Master at least 2 challenge assignments

Late work and resubmissions

- You get a bank of **5** extension days. You can use 1--2 days on any assignment, exam, or project.
- No other late work will be accepted (except in extenuating circumstances!)
- "Not yet mastered" challenge questions, exams, and projects may be resubmitted once

Recap: three ways of fitting linear regression models

- Minimize SSE, via derivatives of ← not appropriate, b/c we don't want $Y - \hat{Y}$
  $$\sum_{i=1}^{n} (Y_i - \underbrace{\beta_0 - \beta_1 X_{i,1} - \cdots - \beta_k X_{i,k}}_{\text{log odds}})^2$$
- Minimize $||Y - \widehat{Y}||$ (equivalent to minimizing SSE)
- Maximize likelihood (for *normal* data, equivalent to minimizing SSE) appropriate, but change distribution

> Which of these three methods, if any, is appropriate for fitting a logistic regression model? Do any changes need to be made for the logistic regression setting?
>
> Discuss with your neighbor for 2--3 minutes.

Maximum likelihood for logistic regression

$$Y_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$$

Suppose we observe independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$. Write down the likelihood function

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$L(\beta) = \prod_{i=1}^{n} f(Y_i; \beta)$$

for the logistic regression problem. Take 2--3 minutes, then we will discuss as a class.

Maximum likelihood for logistic regression

$$L(\beta) = \prod_{i=1}^{n} f(Y_i;\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

$$= \prod_{i=1}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}}\right)^{Y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in}}}\right)^{1-Y_i}\right.$$

> I want to choose β to maximize $L(\beta)$. What are the usual steps to take?

1) Take log: $\ell(\beta) = \log L(\beta)$

2) $\frac{\partial \ell(\beta)}{\partial \beta_0} \overset{set}{=} 0, \quad \frac{\partial \ell(\beta)}{\partial \beta_1} \overset{set}{=} 0, \quad \ldots, \quad \frac{\partial \ell(\beta)}{\partial \beta_n} \overset{set}{=} 0$

Initial attempt at maximizing likelihood

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$L(\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

$$\ell(\beta) = \sum_{i=1}^{n} \left\{ Y_i \log p_i + (1-Y_i)\log(1-p_i) \right\}$$

$$= \sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right\}$$

$$= \sum_{i=1}^{n} \left\{ Y_i(\beta_0+\beta_1 x_i) - \log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right\}$$

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left\{ Y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right\} \overset{\text{set}}{=} 0$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) x_i \overset{\text{set}}{=} 0$$

hard!

Iterative methods for maximizing likelihood

Idea:

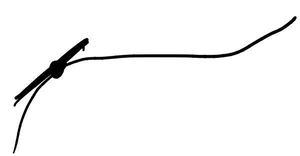
1) Start w/ initial guess $\beta^{(0)}$
2) Update to $\beta^{(1)}$, which is closer to the solution
3) Iterate!

What do we iterate?

Motivation: $U(\beta) = \dfrac{\partial \ell(\beta)}{\partial \beta} = \begin{bmatrix} \dfrac{\partial \ell(\beta)}{\partial \beta_0} \\ \vdots \\ \dfrac{\partial \ell(\beta)}{\partial \beta_K} \end{bmatrix} = \nabla \ell(\beta)$ (gradient)

score function

want to find β^* such that $U(\beta^*) = 0$



Want β^* st $u(\beta^*) = 0$, guess $\beta^{(0)}$

First order Taylor expansion around $\beta^{(0)}$

$$\underbrace{u(\beta^*)}_{= 0} \approx u(\beta^{(0)}) + \frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} (\beta^* - \beta^{(0)})$$

$$\Rightarrow u(\beta^{(0)}) + \frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} (\beta^* - \beta^{(0)}) \approx 0$$

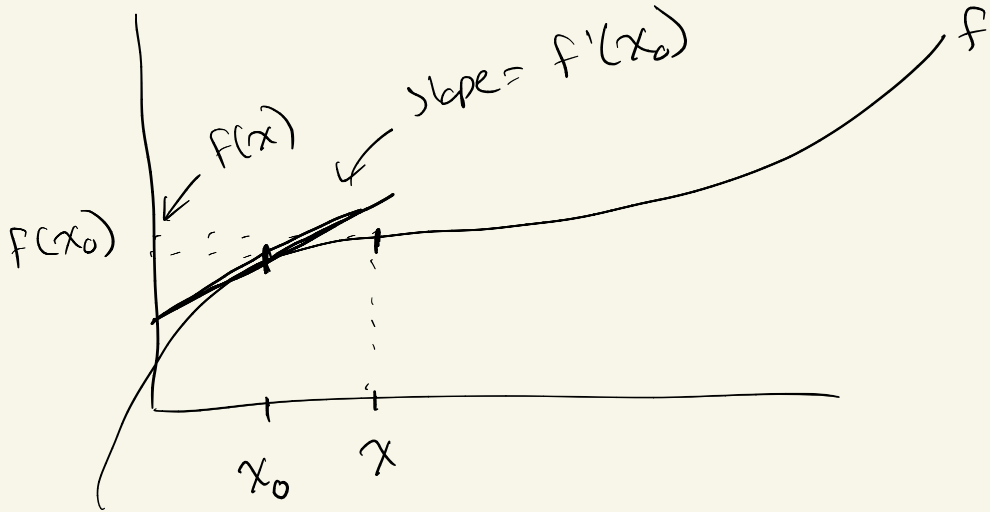
$$\Rightarrow \beta^* \approx \beta^{(0)} - \left(\frac{\partial u(\beta^{(0)})}{\partial \beta^{(0)}} \right)^{-1} u(\beta^{(0)})$$

<u>Iterative procedure</u>:

1) Initial guess $\beta^{(0)}$
2) Update: $r \to r+1$: $\beta^{(r+1)} = \beta^{(r)} - \left(\frac{\partial u(\beta^{(r)})}{\partial \beta^{(r)}} \right)^{-1} u(\beta^{(r)})$
3) Stop when $\beta^{(r)} \approx \beta^{(r+1)}$ (you get to define how close)

Taylor expansion of $f(x)$ around x_0

First order: $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$



$$\frac{f(x) - f(x_0)}{x - x_0} \approx f'(x_0)$$

$$\frac{\partial U(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial^2 \ell(\beta)}{\partial \beta_0^2} & \frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_0 \partial \beta_k} \\ \vdots & & & \\ \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_0} & \cdots & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_k^2} \end{bmatrix}$$

$$\mathcal{J}(\beta) = \frac{-\partial U(\beta)}{\partial (\beta)} \quad \leftarrow \text{observed information matrix}$$

$$\mathcal{I}(\beta) = \mathbb{E}[\mathcal{J}(\beta)] \quad \leftarrow \text{Fisher information matrix}$$

Fisher scoring

Fisher scoring for logistic regression

Practice question: Fisher scoring

Suppose that $\log\left(\dfrac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$, and we have

$$\beta^{(r)} = \begin{bmatrix} -3.1 \\ 0.9 \end{bmatrix}, \qquad U(\beta^{(r)}) = \begin{bmatrix} 9.16 \\ 31.91 \end{bmatrix},$$

$$\mathcal{I}(\beta^{(r)}) = \begin{bmatrix} 17.834 & 53.218 \\ 53.218 & 180.718 \end{bmatrix}$$

> Use the Fisher scoring algorithm to calculate $\beta^{(r+1)}$ (you may use R or a calculator, you do not need to do the matrix arithmetic by hand). Take ~ 5 minutes, then we will discuss.

Alternative to Fisher scoring: gradient ascent

$$Y_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$$

Choose $\beta = (\beta_0, \ldots, \beta_k)^T$ to maximize $L(\beta)$.

Gradient ascent:

Motivation for gradient ascent: walking uphill

Practice question: gradient ascent

Suppose that $\log\left(\dfrac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$, and we have

$$\beta^{(r)} = \begin{bmatrix} -3.1 \\ 0.9 \end{bmatrix}, \qquad U(\beta^{(r)}) = \begin{bmatrix} 9.16 \\ 31.91 \end{bmatrix}$$

- Use gradient ascent with a learning rate (aka step size) of $\gamma = 0.01$ to calculate $\beta^{(r+1)}$.
- The actual maximum likelihood estimate is $\widehat{\beta} = (-3.360, 1.174)$. Does one iteration of gradient ascent or Fisher scoring get us closer to the optimal $\widehat{\beta}$?
- Discuss in pairs for 2--3 minutes.