Logistic regression

Matthew S. Shotwell, Ph.D.

Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN, USA

February 10, 2020

Logistic regression

- models $G|X$ directly
- K classes $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$
- when $K > 2$ called "multinomial logistic regression"
- $P_k = P_k(x, \beta) = Pr(G = \mathcal{G}_k | X = x, \beta)$

Logistic regression

LR model:

- "logit" or log-odds

$$\log\left[\frac{P_k}{P_K}\right] = x\beta_k \quad k = 1, \ldots, K-1$$

- "expit" or "sigmoid" or "logistic"

$$P_k = \frac{\exp(x\beta_k)}{1 + \sum_{l=1}^{K-1} \exp(x\beta_l)}$$

- expit converts $K-1$ numbers to K probabilities that sum to 1
- "sigmoid" used in Keras as output activation

Estimating β_k

- given sample $g_1, \ldots, g_n$, targets $y_1, \ldots, y_n$, inputs $x_1, \ldots, x_n$
- let $\beta = \{\beta_1, \ldots, \beta_K\}$
- minimize average loss in training data

$$\overline{\mathrm{err}}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$

- using cross-entrpoy loss

$$-\sum_{k=1}^{K} y_{ik} \log p_{ik}$$

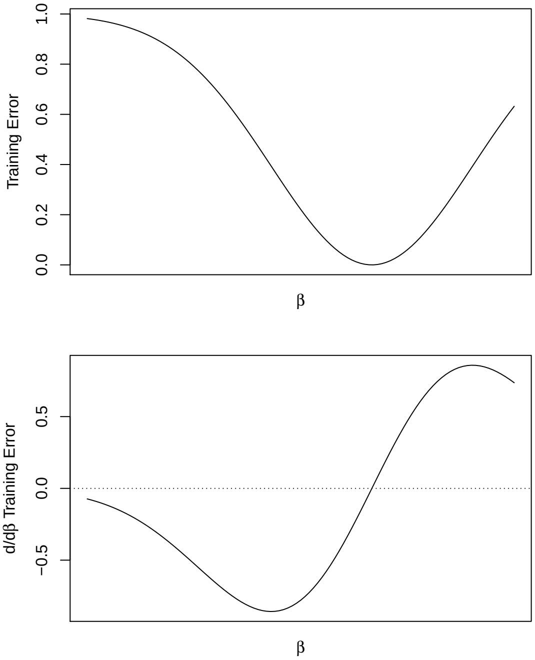
where

$$p_{ik} = P_k(x_i, \beta_k)$$

Estimating β_k

- minimizing the average loss equivalent to maximizing the "log likelihood" function, assuming that outcome has a multinomial distribution:
- log likelihood:

$$l(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log p_{ik}$$

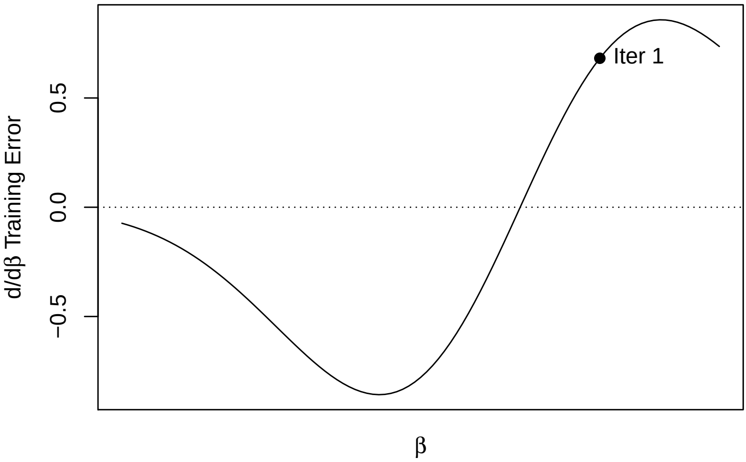


Training Error
0.0
0.2
0.4
0.6
0.8
1.0
β
d/dβ Training Error
−0.5
0.0
0.5
β

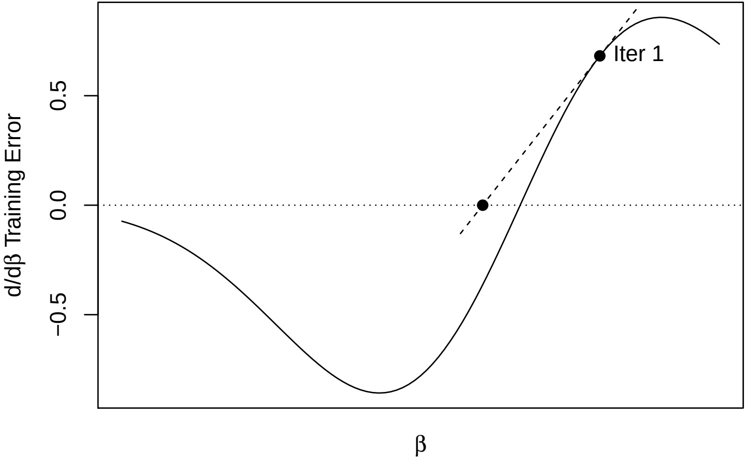
Estimating β_k

- to minimize expected loss, find $\frac{d}{d\beta}\overline{\mathrm{err}}(\beta) = \overline{\mathrm{err}}'(\beta) = 0$
- no closed-form expression for $\overline{\mathrm{err}}'(\beta)$
- need an algorithm to solve
- use Newton-Raphson algorithm

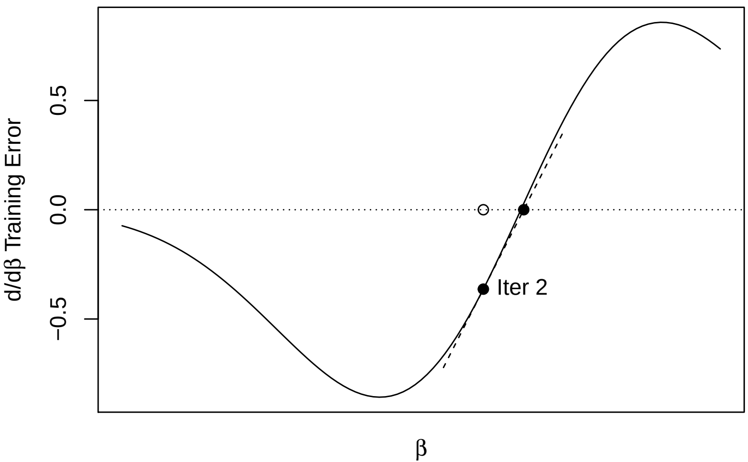
Newton-Raphson algorithm



Newton-Raphson algorithm



Newton-Raphson algorithm



Newton-Raphson algorithm

Use first-order Taylor approximation to linearize $\overline{\mathrm{err}}'$ at starting point β_0

- want to solve $\overline{\mathrm{err}}'(\hat{\beta}) = 0$
- Taylor approximation:

$$\begin{aligned}\overline{\mathrm{err}}'(\hat{\beta}) &\approx \overline{\mathrm{err}}'(\beta_0) + \overline{\mathrm{err}}''(\beta_0)(\beta_0 - \hat{\beta}) \\ \hat{\beta} &\approx \beta_0 - \overline{\mathrm{err}}''(\beta_0)^{-1}\overline{\mathrm{err}}'(\beta_0)\end{aligned}$$

- convert to iterative algorithm:

$$\hat{\beta}_{(m)} = \hat{\beta}_{(m-1)} - \overline{\mathrm{err}}''(\hat{\beta}_{(m-1)})^{-1}\overline{\mathrm{err}}'(\hat{\beta}_{(m-1)})$$

LR vs. LDA

- both express $\log[P_k/P_K]$ as linear in x (see HTF eq. 4.9)
- β estimated differently
- LR makes fewer distributional assumptions
- LR uses cond. prob. $Pr(G|X)$ where $Pr(X)$ ignored
- LDA uses joint prob. $Pr(G, X)$
- LDA smaller $\mathrm{var}(\hat{\beta})$ whem model true (see HTF eq. 4.38)
- LDA can use unclassified observations to help estimate $Pr(X)$
- LR parameters not defined when there is perfect separation
- neither LR nor LDA have natural tuning parameter

Uncertainty in model predictions

- $\hat{G}(x) = \text{argmax}_{\mathcal{G}_k} Pr(G = \mathcal{G}_k | X = x, \hat{\beta})$
- but $\hat{\beta}$ is a sample statistic and therfore has sampling uncertainty given approximately by $N(\hat{\beta}, \hat{I}(\hat{\beta})^{-1})$
- thus $Pr(G = \mathcal{G}_k | X = x, \hat{\beta})$ also has sampling uncertainty
- if using $Pr(G = \mathcal{G}_k | X = x, \hat{\beta})$ to make decisions, might like to know something about this uncertainty

Sampling uncertainty

- statisticians have spent more than 100 years trying to identify the sampling distributions of this and other statistics
- greatest discoveries in statistics were generic strategies for this, e.g., approximate sampling distribution for MLEs, delta method, bootstrap

Sampling distribution for $\hat{\beta}$

- $\hat{\beta}$ is an MLE, thus $\hat{\beta} \rightarrow N(\beta, E_{G|X}[-l''(\beta)]^{-1})$
- approximate $\hat{\beta} \sim N(\hat{\beta}, [-l''(\hat{\beta})]^{-1})$
- Hessian of log likelihood
- Fisher information denoted $I(\beta) = E_{G|X}[-l''(\beta)]$
- observed Fisher information at $\hat{\beta}$ denoted $\hat{I}(\hat{\beta}) = -l''(\hat{\beta})$

Sampling distribution for $Pr(G = \mathcal{G}_k | X = x, \hat{\beta})$

Unfortunately $Pr(G = \mathcal{G}_k | X = x, \hat{\beta})$ is a nonlinear function of $\hat{\beta}$, so can't easily determine sampling distribution. But we can linearize $Pr(G = \mathcal{G}_k | X = x, \hat{\beta})$ in $\hat{\beta}$ using a first-order Taylor approximation:

- let $r(\hat{\beta}) = Pr(G = \mathcal{G}_k | X = x, \hat{\beta})$
- then $r(\hat{\beta}) \approx r(\beta) + r'(\beta)(\hat{\beta} - \beta)$
- thus, since $(\hat{\beta} - \beta) \rightarrow N(0, I(\beta)^{-1})$ it follows approximately that $(r(\hat{\beta}) - r(\beta)) \rightarrow N(0, r'(\beta)^T I(\beta)^{-1} r'(\beta))$
- approximate $r'(\beta)^T I(\beta)^{-1} r'(\beta)$ using $r'(\hat{\beta})^T \hat{I}(\hat{\beta})^{-1} r'(\hat{\beta})$
- this is the "delta method"