

## HW2

$$\beta \rightarrow p \times i$$

①  $\pi_i = P(Y_i=1 | X_i=x_i), \quad Y_i \in \{0,1\}$

$$\log_i + \ln(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{P(Y_i=1 | X_i)}{1-P(Y_i=1 | X_i)}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} = X_i^T \beta$$

$$Y_1, Y_2, \dots, Y_n \sim \text{Bern}(\pi)$$

(PDF is)  $f(y; \pi) = \pi^{y_i} (1-\pi)^{1-y_i}$

$$E[Y_i | X_i] = \pi_i$$

$$\log\left(\frac{E[Y_i | X_i]}{1 - E[Y_i | X_i]}\right) = X_i^T \beta$$

$$e^{X_i^T \beta} = \frac{E[Y_i | X_i]}{1 - E[Y_i | X_i]}$$

$$e^{X_i^T \beta} - E[Y_i | X_i] e^{X_i^T \beta} = E[Y_i | X_i]$$

$$e^{X_i^T \beta} = E[Y_i | X_i] + E[Y_i | X_i] e^{X_i^T \beta}$$

$$e^{X_i^T \beta} = E[Y_i | X_i] (1 + e^{X_i^T \beta})$$

$$\frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} = E[Y_i | X_i] = \pi_i$$

$$L(\beta) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \xrightarrow{\text{sub } \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} \text{ for } \pi_i}$$

$$\prod_{i=1}^n \left( \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} \right)^{y_i} \left( 1 - \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}} \right)^{1-y_i} \downarrow \frac{1}{1 + e^{X_i^T \beta}}$$

$$l(\beta) = \log L(\beta) = \log \prod_{i=1}^n \left( \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right)^{y_i} \left( 1 - \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right)^{1-y_i}$$

$$\sum_{i=1}^n y_i \left( \log \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right) + (1-y_i) \left[ \log \left( 1 - \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right) \right]$$

↙    ↘

$$\sum_{i=1}^n y_i \left[ \log e^{x_i^\top \beta} - \log (1+e^{x_i^\top \beta}) \right]$$

$$\sum_{i=1}^n y_i [x_i^\top \beta - \log(1+e^{x_i^\top \beta})] + (1-y_i) \left[ \log \left( \frac{1+e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} - \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right) \right]$$

$$\sum_{i=1}^n y_i [x_i^\top \beta - \log(1+e^{x_i^\top \beta})] + (1-y_i) \log \frac{1}{1+e^{x_i^\top \beta}}$$

$$\sum_{i=1}^n y_i [x_i^\top \beta - \log(1+e^{x_i^\top \beta})] + (1-y_i) \left[ \log 1 - \log (1+e^{x_i^\top \beta}) \right]$$
~~$$\sum_{i=1}^n y_i x_i^\top \beta - y_i \log(1+e^{x_i^\top \beta}) - \log(1+e^{x_i^\top \beta}) + y_i \log(1+e^{x_i^\top \beta})$$~~

$$l(\beta) : \sum_{i=1}^n y_i x_i^\top \beta - \log(1+e^{x_i^\top \beta})$$



Convex Optimization?

is  $-\log(f)$  convex?  $\rightarrow$  is hessian  $\geq 0$ ?

$$-\log(f) \rightarrow -\sum_{i=1}^n y_i x_i^\top \beta - \log(1+e^{x_i^\top \beta})$$

$$\sum_{i=1}^n \log(1+e^{x_i^\top \beta}) - y_i x_i^\top \beta$$

gradient:  $\frac{\partial l}{\partial \beta} : \sum_{i=1}^n \frac{e^{x_i^\top \beta} x_i}{1+e^{x_i^\top \beta}} - y_i x_i$

$$\text{gradient: } \frac{\partial L}{\partial \beta} = \sum_{i=1}^n \frac{e^{x_i^T \beta} x_i}{1 + e^{x_i^T \beta}} - y_i x_i$$

$$\text{hessian: } \frac{\partial^2 L}{\partial \beta^2} = \sum_{i=1}^n \frac{x_i^T x_i e^{x_i^T \beta} (1 + e^{x_i^T \beta}) - e^{x_i^T \beta} x_i (x_i^T e^{x_i^T \beta})}{(1 + e^{x_i^T \beta})^2}$$

$$x_i^T x_i e^{x_i^T \beta} + x_i^T x_i e^{2x_i^T \beta} - x_i^T x_i e^{2x_i^T \beta}$$

always positive  $\rightarrow \sum \frac{x_i^T x_i e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})^2} \geq 0$ , thus MLE convex optimization

Likelihood, gradient, hessian

$$\text{actual gradient: } l(\beta) = \sum_{i=1}^n y_i x_i^T \beta - \log(1 + e^{x_i^T \beta})$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^n y_i x_i^T - \frac{x_i^T e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \rightarrow \sum_{i=1}^n (y_i - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}) x_i = \sum_{i=1}^n (y_i - \pi_i) x_i$$

$$\frac{\partial^2 l}{\partial \beta^2} = \sum_{i=1}^n 0 - \frac{x_i^T x_i e^{x_i^T \beta} (1 + e^{x_i^T \beta}) - x_i e^{x_i^T \beta} (x_i^T e^{x_i^T \beta})}{(1 + e^{x_i^T \beta})^2} =$$

$$= - \sum_{i=1}^n \frac{x_i^T x_i e^{x_i^T \beta} [(1 + e^{x_i^T \beta}) - e^{x_i^T \beta}]}{(1 + e^{x_i^T \beta})^2} = - \sum_{i=1}^n \frac{x_i^T x_i e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})^2}$$

$$= - \sum_{i=1}^n x_i^T \left( \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right) \left( \frac{1}{1 + e^{x_i^T \beta}} \right) x_i$$

$$= - \sum_{i=1}^n x_i^T (\pi_i (1 - \pi_i)) x_i$$

Remember

$$\pi_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

$$1 - \pi_i = \frac{1}{1 + e^{x_i^T \beta}}$$

## Update

Steepest Descent

$$\rightarrow \beta_{t+1} = \beta_t - \alpha \frac{\partial l}{\partial \beta} \Rightarrow \beta_{t+1} = \beta_t - \alpha \left[ \sum_{i=1}^n (y_i - \pi_i) x_i \right]$$

• Newton

$$\rightarrow \beta_{t+1} = \beta_t - \left[ \frac{\partial^2 l}{\partial \beta^2} \right]^{-1} \frac{\partial l}{\partial \beta} \Rightarrow$$

$$\beta_{t+1} = \beta_t - \frac{\sum_{i=1}^n (y_i - \pi_i) x_i}{-\sum_{i=1}^n x_i^\top (\pi_i(1-\pi_i)) x_i} \quad \text{where} \quad \pi_i = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}}$$

## Extra Credit A:

$$I(\beta) : -E[H(\beta)] : -E \left[ -x_i^\top (\pi_i(1-\pi_i)) x_i \right] = x_i^\top (\pi_i(1-\pi_i)) x_i$$

↑  
not dep. on  $y_i$ /data.

$$I_n(\beta) : -H(\beta) = x_1^\top x_1 (\pi_i(1-\pi_i))$$

## Extra Credit B:

$$\text{Probit Regression: } \phi'(Pr[Y_i=1 | X_i]) = x_i^\top \beta$$

how to uninvert probit..

z-score in essence?

$$Pr[Y_i=1 | X_i] = \phi(x_i^\top \beta)$$

$$Y_i \sim \text{Bern}(\pi_i)$$

$$\text{pdf: } f(y_i; \pi) = \pi^{y_i} (1-\pi)^{1-y_i}$$

$$E[Y_i=1 | X_i] = \phi(x_i^\top \beta) \quad \text{area to left of } x_i^\top \beta ?$$

$$\text{CDF} = \int_{-\infty}^{x_i^\top \beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x^2} dx ?$$

$$\mu=0 \quad \sigma^2=1$$

$$L(\beta) = \prod_{i=1}^n \frac{y_i}{\pi_i} (1 - \pi_i)^{1-y_i} \xrightarrow{\text{sub } \phi(x_i^\top \beta) \text{ for } \pi_i} \prod_{i=1}^n \left( \phi(x_i^\top \beta) \right)^{y_i} \left( 1 - \phi(x_i^\top \beta) \right)^{1-y_i}$$

$$\phi(x_i^\top \beta) = \int_{-\infty}^{x_i^\top \beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

$$L(\beta) = \sum_{i=1}^n \left[ y_i \log(\phi(x_i^\top \beta)) + (1-y_i) \log(1 - \phi(x_i^\top \beta)) \right]$$

$$\frac{\partial L}{\partial \beta} :$$

$$\begin{aligned} \frac{\partial L}{\partial \beta} & [y_i \log(\phi(x_i^\top \beta))] \\ &= \left( y_i \cdot \frac{1}{\phi(x_i^\top \beta)} \cdot \frac{\partial}{\partial \beta} \left( \int_{-\infty}^{x_i^\top \beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right) \right) + (\log(\phi(x_i^\top \beta)) \cdot 0) \\ &= \frac{y_i}{\phi(x_i^\top \beta)} \cdot \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \cdot x_i \right) \end{aligned}$$

chain rule

$$\frac{\partial L}{\partial \beta} [(1-y_i) \log(1 - \phi(x_i^\top \beta))]$$

$$(1-y_i) \cdot -\frac{1}{1-\phi(x_i^\top \beta)} \cdot \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \cdot x_i \right) + (\log(1 - \phi(x_i^\top \beta))) \cdot 0$$

entire gradient:

$$\sum_{i=1}^n \frac{y_i}{\phi(x_i^\top \beta)} \cdot \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right) x_i - \frac{1-y_i}{1-\phi(x_i^\top \beta)} \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right) x_i$$

$$\sum_{i=1}^n x_i^\top \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \left[ \frac{y_i}{\phi(x_i^\top \beta)} - \frac{1-y_i}{1-\phi(x_i^\top \beta)} \right] \quad \text{prod. rule}$$

i think i will know M

$$\text{hessian: } \frac{\partial^2 L}{\partial \beta^2} = \sum_{i=1}^n \left( x_i^\top \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right) \left[ \frac{\partial}{\partial \beta} \left( \frac{y_i}{\phi(x_i^\top \beta)} - \frac{1-y_i}{1-\phi(x_i^\top \beta)} \right) \right] +$$

$$\left[ \frac{y_i}{\phi(x_i^\top \beta)} - \frac{1-y_i}{1-\phi(x_i^\top \beta)} \right] \left( x_i^\top \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \cdot -x_i^\top \beta x_i \right)$$

First  
half:

$$\left( x_i^\top \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right) \left[ \frac{\partial}{\partial \beta} \left( \frac{y_i}{\Phi(x_i^\top \beta)} - \frac{1-y_i}{1-\Phi(x_i^\top \beta)} \right) \right]$$

$$\begin{aligned} & \text{quotient rule:} \\ & \frac{\Phi(x_i^\top \beta)(0) - y_i \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right)}{[\Phi(x_i^\top \beta)]^2} = -y_i \frac{\left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right)}{[\Phi(x_i^\top \beta)]^2} \\ & \frac{(1-\Phi(x_i^\top \beta))(0) - (1-y_i) \left( -\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right)}{(1-\Phi(x_i^\top \beta))^2} = \frac{(1-y_i) \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right)}{(1-\Phi(x_i^\top \beta))^2} \end{aligned}$$

Full hessian:

$$\sum_{i=1}^n -y_i \frac{\left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right)}{[\Phi(x_i^\top \beta)]^2} + \frac{(1-y_i) \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \right)}{(1-\Phi(x_i^\top \beta))^2} + \\ \left[ \frac{y_i}{\Phi(x_i^\top \beta)} - \frac{1-y_i}{1-\Phi(x_i^\top \beta)} \right] \left( x_i^\top \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^\top \beta)^2}{2}} \cdot -x_i^\top \beta x_i \right)$$

↳ can see observed & expected info will be diff. b/c  
y<sub>i</sub> present in hessian

## Problem 4

### Example 2: Censored Exponential Data

- Suppose we have survival times  $t_1, \dots, t_n \sim \text{Exp}(\lambda)$
- Do not observe all survival times b/c some are censored at times  $c_1, \dots, c_h$
- Actually observe  $y_1, \dots, y_n$  where  $y_i = \min(t_i, c_i)$ 
  - Also have indicator  $\delta_i$  where  $\delta_i = 1$  if  $t_i \leq c_i$
  - i.e.  $\delta_i = 1$  if not censored and  $\delta_i = 0$  if censored

$$= p(t_i | \theta)$$

What is  $p(y_i | z_i | \theta)$ , the complete data density

- What is  $z$ ? → the missing data? → can't be label like in gaussian mixture model b/c not mixing 2 distributions

\*  $z$  is survival time of those who were censored

Exponential pdt  $\frac{1}{\lambda} e^{-\frac{t}{\lambda}}$ ,  $E(X) = \lambda$

$$p(y_i | z_i | \theta) = p(t_i | \theta) \quad t_i = \delta_i y_i + (1 - \delta_i) z_i$$

↳  $E(t_i)$  for E-step

$$p(t_i | \lambda) = \frac{1}{\lambda} e^{-\frac{t_i}{\lambda}}$$

$$p(y_i, t_i | \theta) = \frac{1}{\lambda} e^{-(\delta_i y_i + (1 - \delta_i) z_i) / \lambda} \rightarrow \text{complete data density}$$

$$\log p(y_i, t_i | \theta) = \sum_i^n \log(\lambda)^{-1} - \frac{1}{\lambda} (\delta_i y_i + (1 - \delta_i) z_i) \\ - n \log \lambda - \frac{1}{\lambda} \sum_i (\delta_i y_i + (1 - \delta_i) z_i)$$

From  
Lab

$$Q(\lambda | \lambda_0) = E_{z_i} \left[ -n \log \lambda - \frac{1}{\lambda} \sum_i (\delta_i y_i + (1 - \delta_i) z_i) \right]$$

$$= -n \log \lambda - \frac{1}{\lambda} \sum_i \left[ E_z (\delta_i y_i + (1 - \delta_i) z_i) \right]$$

$$= -n \log \lambda - \frac{1}{\lambda} \sum_i \left[ E_{\delta_i=1} (\delta_i y_i) + E_{\delta_i=0} ((1 - \delta_i) z_i) \right]$$

both  $y_i$  and  $\delta_i$

$\delta_i$  known/observed already

Need  $E(z_i | y_i, \lambda)$

$E(z_i | z_i > y_i) = y_i + E(y_i) = y_i + \lambda$   
↳ conditional expectation w/  
memorylessness of exponential dist  
 $E(X | X > t) = t + E(X)$

Now minimize  $\Theta$  by deriving wrt  $\lambda$  and setting to 0

$$\Theta(\lambda | \lambda_0) = -n \log \lambda - \frac{1}{\lambda} \sum_i [E(\delta_i y_i) + E[(1-\delta_i)z_i]]$$

$$= -n \log \lambda - \frac{1}{\lambda} \sum_i [\delta_i y_i + (1-\delta_i)(y_i + \lambda)]$$

$$= -n \log \lambda - \frac{1}{\lambda} \sum_i [\cancel{\delta_i y_i} + y_i + \lambda - \cancel{\delta_i y_i} - \delta_i \lambda]$$

$$= -n \log \lambda - \frac{1}{\lambda} \sum_i [y_i + (1-\delta_i)\lambda]$$

$$= -n \log \lambda - \frac{1}{\lambda} [\sum_i y_i + \lambda \sum_i (1-\delta_i)] \quad ?$$

$$= -n \log \lambda - \frac{1}{\lambda} \sum_i y_i - \sum_i (1-\delta_i)$$

$$\frac{\partial \Theta}{\partial \lambda} = -\frac{n}{\lambda} + \frac{\sum y_i}{\lambda^2} = 0$$

$$\frac{\sum y_i}{\lambda^2} = \frac{n}{\lambda}$$

$$\sum y_i = n\lambda$$

$$\text{M step: } \frac{\sum y_i}{n} = \lambda \quad \checkmark$$

↳ use bootstrap for CI