HW2

$\beta \to p \times i$

1. $\pi_i = P(Y_i = 1 \mid X_i = x_i)$, $Y_i \in (0,1)$

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{P(Y_i=1 \mid X_i)}{1-P(Y_i=1 \mid X_i)}\right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi} = X_i^T \beta$$

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

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

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

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

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

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

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

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

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

sub $\frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}$ for π_i

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \longrightarrow$$

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

$$\downarrow \frac{1}{1+e^{X_i^T\beta}}$$

$$(e^{x_i^T\beta})^{y_i}-(1+e^{x_i^T\beta})$$

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

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

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

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

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

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

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

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

$\downarrow$

<u>Convex Optimization?</u>

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

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

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

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

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

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

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

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 this convex optimization

---

Likelihood, gradient, hessian

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

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

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

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

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

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

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

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

- Newton

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

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

Extra Credit A:

$$I(\beta) = -E[H(\beta)] = -E[-(X_i^T(\pi_i(1-\pi_i)X_i)] = X_i^T (\pi_i(1-\pi_i)) X_i$$

↑ not dep. on y_i/data!

$$I_n(\beta) = -H(\beta) = X_i^T X_i (\pi_i (1-\pi_i))$$

Extra Credit B:

Probit Regression! $\Phi^{-1}(Pr[Y_i = 1 \mid X_i]) = X_i^T \beta$

z-score in essence?

how to uninvert probit..

$$Pr[Y_i = 1 \mid X_i] = \Phi(X_i^T \beta)$$

$Y_i \sim Bern(\pi_i)$

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

$$E[Y_i = 1 \mid X_i] = \Phi(X_i^T \beta)$$

area to left of $X_i^T \beta$?

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

$$CDF = \int_{-\infty}^{X_i^T \beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \;?$$

sub $\Phi(x_i^T\beta)$ for π_i

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \longrightarrow$$

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

$$\prod_{i=1}^{n} \left(\Phi(x_i^T\beta)\right)^{y_i} \left(1 - \Phi(x_i^T\beta)\right)^{1-y_i}$$

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

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

$$\frac{\partial \ell}{\partial \beta} \left[y_i \log(\Phi(x_i^T\beta)) \right]$$

$$= \left(y_i \cdot \frac{1}{\Phi(x_i^T\beta)} \cdot \frac{\partial}{\partial \beta}\left(\int_{-\infty}^{x_i^T\beta} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \right)\right) + \left(\log(\Phi(x_i^T\beta)) \cdot 0\right)$$

$$= \frac{y_i}{\Phi(x_i^T\beta)} \cdot \left(\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^T\beta)^2}{2}}\right) \cdot \underbrace{x_i}_{\text{chain rule}} \right)$$

$$\frac{\partial}{\partial \beta} \left[(1-y_i)(\log(1-\Phi(x_i^T\beta))) \right]$$

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

entire gradient:
$$\sum_{i=1}^{n} \frac{y_i}{\Phi(x_i^T\beta)} \cdot \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^T\beta)^2}{2}} \right) x_i - \frac{1-y_i}{1-\Phi(x_i^T\beta)} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^T\beta)^2}{2}} \right) x_i$$

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

prod. rule

i think i will kms ♡

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

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

First half:

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

quotient rule:

$$\frac{\Phi(x_i^T\beta)(0) - y_i\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^T\beta)^2}{2}}\right)}{[\Phi(x_i^T\beta)]^2} = \frac{-y_i\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x^T\beta)^2}{2}}\right)}{[\Phi(x_i^T\beta)]^2}$$

$$\frac{(1-\Phi(x_i^T\beta)(0) - (1-y_i)\left(-\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^T\beta)^2}{2}}\right)}{(1-\Phi(x_i^T\beta))^2} = \frac{(1-y_i)\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i^T\beta)^2}{2}}\right)}{(1-\Phi(x_i^T\beta))^2}$$

Full hessian:

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

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

↳ can see observed & expected info will be diff. b/c y_i present in hessian

Problem 4

Example 2: Censored Exponential Data

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

What is $p(y, z \mid \theta) = p(t, \lambda)$, 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 pdf: $\frac{1}{\lambda} e^{-\frac{x}{\lambda}}$, $E(X) = \lambda$

$p(y, z \mid \theta) = p(t \mid \theta)$ $\quad t \sim Exp(\lambda)$ $\quad t_i = \delta_i y_i + (1 - \delta_i) z_i$

$$p(t \mid \lambda) = \frac{1}{\lambda} e^{-t/\lambda}$$

$$p(y, t \mid \theta) = \frac{1}{\lambda} e^{-(\delta_i y_i + (1-\delta_i) z_i)/\lambda}$$ → complete data density

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

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

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

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

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

From Lab

both y_i and δ_i known/observed already

Need $E(z_i \mid y, \lambda)$

$E(z_i \mid z_i > y_i) = y_i + E(y_i) = y_i + \lambda$

↳ conditional expectation w/ memorylessness of exponential dist

$E(X \mid X > t) = t + E(X)$

Now maximize Q by deriving wrt λ and setting to 0

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

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

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

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

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

$$0 = -\frac{n}{\lambda} + \frac{\sum y_i + \lambda\sum_i(1-\delta_i)}{\lambda^2}$$

$$\frac{n}{\lambda} = \frac{\sum y_i + \lambda(n-\sum\delta_i)}{\lambda^2}$$

$$n\lambda = \sum y_i + \lambda(n-\sum\delta_i)$$

$$n\cancel{\lambda} - \cancel{\lambda}n + \lambda\sum\delta_i = \sum y_i$$

$$\lambda\sum\delta_i = \sum y_i$$

$$\lambda = \frac{\sum y_i}{\sum\delta_i} \quad - \quad \underline{\text{M step}}$$

Use Louis or Bootstrap for CI

↓

na derivation

how to implement in R...