HW2

2025-02-12

TODO:

Put 2A and 2C in on Latex implement MM (uh oh) # Setup

1

Let $\pi_i = P(Y_i = 1 \mid X_i = x_i), Y_i \in \{0,1\}$. The logistic regression model is given by:

$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{P(Y_i=1|X_i)}{1-P(Y_i=1|X_i)}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} = X_i^T\beta$

where β is a $p \times 1$ vector.

The response variables follow a Bernoulli distribution: $Y_1, Y_2, \ldots, Y_n \sim \text{Bern}(\pi)$, with probability mass function:

$f(y_i \mid \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

The expectation of Y_i given X_i is $E[Y_i \mid X_i] = \pi_i$. Since $\log\left(\frac{E[Y_i|X_i]}{1-E[Y_i|X_i]}\right) = X_i^T\beta$, we exponentiate both sides:

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

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

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

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

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

Likelihood

The likelihood function is:

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

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

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

The log-likelihood function is given by:

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

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

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

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

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

Thus, the final form of the log-likelihood function is:

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

Gradient

$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} y_i X_i^T - \sum_{i=1}^{n} \frac{X_i e^{X_i^T \beta}}{1+e^{X_i^T \beta}}$

Factoring common terms out: $\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \frac{e^{X_i^T \beta}}{1+e^{X_i^T \beta}} \right) X_i$

Rewriting in terms of π_i we get:

$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} (y_i - \pi_i) X_i$

Hessian

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

Simplifying: $\frac{\partial^2 \ell}{\partial \beta^2} = -\sum_{i=1}^{n} \frac{X_i^T X_i e^{X_i^T \beta}[(1+e^{X_i^T \beta}) - e^{X_i^T \beta}]}{(1+e^{X_i^T \beta})^2}$

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

Since:

$\pi_i (1 - \pi_i) = \frac{e^{X_i^T \beta}}{(1+e^{X_i^T \beta})^2}$

$\frac{\partial^2 \ell}{\partial \beta^2} = -\sum_{i=1}^{n} X_i^T (\pi_i (1 - \pi_i)) X_i$

Newton's Method Update

$\beta_{t+1} = \beta_t - \left(\frac{\partial^2 \ell}{\partial \beta^2} \right)^{-1} \frac{\partial \ell}{\partial \beta}$

$\beta_{t+1} = \beta_t - \left(\frac{\sum_{i=1}^{n} (y_i - \pi_i) X_i}{-\sum_{i=1}^{n} X_i^T (\pi_i (1-\pi_i)) X_i} \right)$

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

Convex Optimization

Gradient of -log f is the negative Hessian:

$-\frac{\partial^2 \ell}{\partial \beta^2} = +\sum_{i=1}^{n} X_i^T (\pi_i (1 - \pi_i)) X_i \geq 0$

The negative hessian is always positive, thus this is a convex optimization problem.

2

Add in Latex here

3

Newton

```
## [1] 0.1585398 0.1637085
```

```
##           [,1]
## [1,] 0.9189634
## [2,] 0.2487389
```

GLM

```
## (Intercept)           x          xx
##   0.9189634          NA   0.2487389
```

4

Derivation of EM algorithm

- Survival times $t_1, \ldots, t_n \sim \text{Exp}(\lambda)$.
- Times censored at $c_1, \ldots, c_n$.
- We observe y_i, where $y_i = \min(t_i, c_i)$.
- Indicator $\delta_i = 1$ if $t_i \leq c_i$ (not censored), $\delta_i = 0$ if $t_i > c_i$ (censored).

We need $p(y \mid z, \theta)$ the complete data density.

z represents the survival times of those who were censored (the missing data in this case)

The exponential density function is:

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

where $E(\lambda) = \lambda$.

Given that $t \sim \text{Exp}(\lambda)$:

$t_i = \delta_i y_i + (1 - \delta_i) z_i$

(We need the expected value of this for the E step)

and:

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

Thus, the complete data density is:

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

Log-Likelihood Function

Taking the log:

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

$$-n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} (\delta_i y_i + (1-\delta_i) z_i)$$

Q function:

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

Since y_i and δ_i are already observed:

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

For censored observations ($\delta_i = 0$), we use the memorylessness property of exponential distribution to get:
$E[z_i \mid z_i > y_i, \lambda] = y_i + \frac{1}{\lambda}$
Thus:
$Q(\lambda \mid \lambda_0) = -n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} \left(\delta_i y_i + (1-\delta_i)(y_i + \lambda)] \right)$
Expanding:
$Q(\lambda \mid \lambda_0) = -n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} \left(\delta_i y_i + y_i + \lambda - \delta_i y_i - \delta_i \lambda)] \right)$
Simplifying:
$Q(\lambda \mid \lambda_0) = -n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} \left(\delta_i y_i + y_i + \lambda - \delta_i y_i - \delta_i \lambda)] \right)$
Factoring:

$$Q(\lambda \mid \lambda_0) = -n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} (y_i + (1-\delta_i)\lambda)$$

$$-n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} (1-\delta_i)$$

Taking the derivative of Q w.r.t λ and setting it to 0:

$$\frac{\partial \ell}{\partial \lambda} = \frac{-n}{\lambda} + \frac{\sum_{i=1}^{n} y_i}{\lambda^2}$$

$$\frac{\sum_{i=1}^{n} y_i}{\lambda^2} = \frac{n}{\lambda}$$

$$\sum_{i=1}^{n} y_i = n\lambda$$

M-step: $\frac{\sum_{i=1}^{n} y_i}{n} = \lambda$

Implementation

```
## Estimated lambda from EM: 130.1797

## 95% CI for lambda (EM): [ 101.8725 , 158.4869 ]

## Call:
## phreg(formula = Surv(time, status) ~ 1, data = veteran, dist = "weibull",
##     shape = 1)
##
## Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
## log(scale)                    4.869               0.088     0.000
##
##  Shape is fixed at  1
##
## Events                     128
## Total time at risk         16663
## Max. log. likelihood      -751.22

## Estimated lambda from phreg: 130.1797

## 95% CI for lambda (phreg): [ 109.4726 , 154.8035 ]
```

The fitted λ obtained using EM algorithm was 130.1797, with a CI of [101.8725 , 158.4869]. I monitered convergence by comparing the difference between the observed likelihood at the current iteration and the observed likelihood at the previous iteration - if the difference was $<$ 1e-12, the algorithm was determined to have "converged". The variance for the confidence interval was computed using bootstrap sampling.

Using the phreg command, the estimated lambda was 130.1797 with a CI of [109.4726 , 154.8035]. Using our EM algorithm, we converged to the same estimated lambda/mean survival time, but it looks like our EM confidence interval was slightly wider.

Extra Credit A

Fisher/expected information:

$$I(\beta) = -E[H(\beta)]$$

Expanding,

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

Since expectation does not depend on the observed data:

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

Observed information:

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

Thus, observed and expected information are the same for logistic regression.

Extra Credit B

Probit regression

$$\Phi^{-1}\left(P(Y_i = 1 \mid X_i)\right) = X_i^T \beta$$
$$Pr[Y_i \mid X_i] - \Phi(X_i^T \beta)$$
$$Y_i \sim Bern(\pi_i)$$

$$Y_i \sim Bern(\pi_i)$$

$$f(y_i \mid \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

CDF for Φ :

$$\Phi(X_i^T \beta) = \int_{-\infty}^{X_i^T \beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$

Log-Likelihood Function

The likelihood function is:

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

Substituting $\pi_i = \Phi(X_i^T \beta)$:

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

Taking the log:

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

Gradient Calculation

We start with the left half first:

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

Chain rule:

$$\frac{\partial}{\partial \beta}[y_i \log(\Phi(X_i^T \beta))] = \left(\frac{y_i}{\Phi(X_i^T \beta)}\right) \cdot \frac{\partial}{\partial \beta}\left(\int_{-\infty}^{X_i^T \beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx\right)$$

Since the derivative of the a CDF is the PDF:

$$\frac{\partial}{\partial \beta}\Phi(X_i^T\beta) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(X_i^T\beta)^2}{2}}X_i$$

$$\frac{y_i}{\Phi(X_i^T\beta)} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{(X_i^T\beta)^2}{2}}X_i$$

Now for the right term:

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

Applying the chain rule:

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

Full Gradient

Putting it all together:

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

Factoring:

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

Hessian Calculation

Taking the second derivative using the product rule:

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

The first term in the Hessian calculation:

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

Using the quotient rule:

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

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

Similarly, for the second term:

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

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

Full Hessian

All together:

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

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

We can see that the expected and observed information will be different because y_i is present in the Hessian, so it's dependent on our data.