BIOS 731 Advanced Statistical Computing
Fall 2020
Homework 1

Due 9/18/2020 Friday at 11:59pm

Instruction:

- Please submit both write-ups and programs in two separate files.
- All submissions need to be in electronic format.
- The write-up is preferred be in pdf format (written in Word or LaTeX, or scanned hand-written document). You can take a picture of hand-written document and submit JPG if you don't have a scanner, but make sure the picture is clear and readable. Name the file **BIOS731_NAME_hw1.EXT**. Replace NAME by your name, and EXT by proper extension name (pdf or jpg).
- The programs need to be written in a high-level language (no compilation required), and R is highly recommended. The codes for all problems need to be saved in a **single** file named **BIOS731_NAME_hw1.EXT**. Replace NAME by your name, and EXT by proper extension name, e.g., R, sas, m, py, etc. Provide adequate comments in the codes to clearly mark the section for different questions. The codes should generate all results and figures in the homework. Please make sure the codes are "self-contained", e.g., does not depend on platform, can be run at any other machine in any subdirectory, and does not require user input.
- Total is 100 points, with 20 points bonus for the last question. Partial credit will be given.

Problem 1: Permutation test and bootstrap in linear regression. (20 pts)

Consider a multiple linear regression model $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i$. Assume $b_0 = 10$, $b_1 = 1$, $b_2 = 0.5$.

1. Assume $\epsilon_i \sim N(0, 1)$, use permutation test to test null hypothesis $H_0 : b_2 = 0$. Report result p-value, and then compare with the one from using R (from the `lm` function). You can use following codes to simulate data (y_i, x_{1i} and x_{2i}).

```
n=100
b0=10; b1=1; b2=0.5
x1 = rnorm(n); x2=rnorm(n); eps=rnorm(n)
y = b0+b1*x1+b2*x2+eps
```

2. Use non-parametric bootstrap to estimate the 95% confidence interval (CI) for $\hat{b}_2$. Generate data using different residual distributions: (1) $\epsilon_i \sim N(0,1)$; (2) $\epsilon_i \sim t(5)$. Compare the estimated CIs to theoretical values assuming normally distributed residuals (using R `confint` function). Comment on the results.

Problem 2: Iteratively reweighted least squares (10 pts)

Implement the Iteratively reweighted least squares (IRLS) algorithm for Poisson generalized linear model (GLM) with canonical link (log). You should write a function named "`poisreg`", which takes a response `y` (a vector of integers) and a covariate matrix `X`. Optionally, the function takes the maximum number of iterations allowed, and a scale for tolerance parameter to check convergence. The function should return the estimated coefficient, and the estimated variance/covariance matrix of the estimates.

Compare your results with that from the `glm` function in R. You can use the following codes to simulate data and run `glm`:

```
n = 100 ## number of observations
p = 3 ## number of covariates
## generate X, the covariates
X = cbind(1, matrix(rnorm(n*p), ncol=p))
beta = c(1, .5, 1, 2)
mu = exp(X %*% beta)
## generate Y, the outcome
Y = rpois(n, mu)
### use R's glm function to fit
fit = glm(Y~X-1, family=poisson)
coef(fit) ## estimated coefficients
vcov(fit) ## estimated variance/covariance matrix of the estimates
```

Problem 3: EM algorithm in two-component mixture model (10 pts)

Use EM algorithm to fit a two-component Poisson mixture model. You need to provide derivations of the E- and M-steps. Then mimic the two-component Normal mixture model example in the EM class to write a function to fit a two-component Poisson mixture model. You should name your function `twoPois`. It takes a vector `y` and reports the estimated mixing proportion and two Poisson rates. You can use the following codes to simulate data:

```
Y = c(rpois(30, lambda=5), rpois(70, lambda=15))
```

Problem 4: EM algorithm in censored data (30 pts)

The standard linear regression model can be written as

$$Y_i = X_i^{\mathrm{T}}\beta + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Suppose that X_i is observed, but that rather than observing Y_i, we observe $Y_i^* = \min\{Y_i, c\}$, where the censoring value c is known and is constant for all i. The EM algorithm offers a vehicle for estimating the parameters (θ, σ^2) in the presence of censoring. Suppose we reorder the data so that the first m subjects have observed Y_i ($Y_i^* = Y_i$) and the rest $n - m$ subjects have censored Y_i ($Y_i^* = c$).

1. Write out the complete-data log likelihood. Identify the conditional expectations to be evaluated in the E step.

2. Show that

$$\begin{aligned} \mathrm{E}(Y_i | Y_i^* = c, \beta^{(k)}, \sigma^{(k)}) =& X_i^{\mathrm{T}}\beta^{(k)} + \sigma^{(k)}\psi\left(\frac{c - X_i^{\mathrm{T}}\beta^{(k)}}{\sigma^{(k)}}\right) \\ \mathrm{E}[(Y_i - X_i^{\mathrm{T}}\beta^{(k)})^2 | Y_i^* = c, \beta^{(k)}, \sigma^{(k)}] =& \sigma^{2(k)}\left\{1 + \frac{c - X_i^{\mathrm{T}}\beta^{(k)}}{\sigma^{(k)}}\psi\left(\frac{c - X_i^{\mathrm{T}}\beta^{(k)}}{\sigma^{(k)}}\right)\right\}, \end{aligned}$$

   where $\psi(z) = \phi(z)/(1 - \Phi(z))$, and $\phi(z)$ and $\Phi(z)$ are corresponding pdf and cdf of a standard normal random variable.

   Hint:

$$\begin{aligned} &\int_u^\infty z\phi(z)dz = (2\pi)^{-1/2}\int_u^\infty \exp\left(-\frac{z^2}{2}\right)d\frac{z^2}{2} = (2\pi)^{-1/2}\exp(-\frac{u^2}{2}) = \phi(u); \\ &\int_u^\infty z^2\phi(z)dz = -(2\pi)^{-1/2}\int_u^\infty z d\exp\left(-\frac{z^2}{2}\right) = (2\pi)^{-1/2}u\exp\left(-\frac{u^2}{2}\right) + (2\pi)^{-1/2}\int_u^\infty \exp\left(-\frac{z^2}{2}\right)dz \\ &\qquad = u\phi(u) + 1 - \Phi(u) \end{aligned}$$

3. Write out the formulas for updating parameters in the M step.

Problem 5: MM algorithm in logistic regression (30 pts)

In the logistic regression model of Example 5 (lecture "MM Algorithm"), it is possible to separate parameters and avoid matrix inversion altogether.

1. In constructing a minorizing function, first prove the inequality

$$-\log\left\{1+\exp(X_i^{\mathrm{T}}\theta)\right\} \geq -\log\left\{1+\exp(X_i^{\mathrm{T}}\theta^{(k)})\right\} - \frac{\exp(X_i^{\mathrm{T}}\theta)-\exp(X_i^{\mathrm{T}}\theta^{(k)})}{1+\exp(X_i^{\mathrm{T}}\theta^{(k)})},$$

with equality when $\theta = \theta^{(k)}$. This eliminates the log terms.

2. Now apply the arithmetic-geometric mean inequality to the exponential function $\exp(X_i^{\mathrm{T}}\theta)$ to separate parameters. Assuming that θ has p components and that there are n observations, show that these maneuvers lead to the minorizing function

$$g(\theta|\theta^{(k)}) = -\frac{1}{p}\sum_{i=1}^{n}\frac{\exp(X_i^{\mathrm{T}}\theta^{(k)})}{1+\exp(X_i^{\mathrm{T}}\theta^{(k)})}\sum_{j=1}^{p}\exp\{pX_{ij}(\theta_j-\theta_j^{(k)})\} + \sum_{i=1}^{n}Y_iX_i^{\mathrm{T}}\theta$$

up to a constant that does not depend on θ.

3. Finally, prove that maximizing $g(\theta|\theta^{(k)})$ consists in solving the equation

$$-\sum_{i=1}^{n}\frac{\exp(X_i^{\mathrm{T}}\theta^{(k)})X_{ij}\exp(-pX_{ij}\theta_j^{(k)})}{1+\exp(X_i^{\mathrm{T}}\theta^{(k)})}\exp(pX_{ij}\theta_j) + \sum_{i=1}^{n}Y_iX_{ij} = 0$$

for each j. This can be accomplished numerically and you do not need to show that.

Problem 6: Compare different algorithms for logistic regression (bonus 20 pts)

For the logistic regression model example, (Example 5 in lecture "MM Algorithm"):

1. Outline the (standard) Newton-Raphson algorithm and the Fisher Scoring algorithm. Show the relation between Newton-Raphson and Fisher Scoring.

2. Implement the Newton-Raphson algorithm in R.

3. Implement the MM algorithm of Example 5 in R.

4. Conduct a simulation study. For 1,000 individuals, generate the binary response from

$$\Pr(Y_i = 1) = \frac{\exp(X_i\theta)}{1 + \exp(X_i\theta)},$$

where $\theta = 0.3$ and $X_i \sim N(0, 1)$. Apply your NR and MM algorithms to this data set; select your own starting value and stopping criterion, but make sure they are the same for the two algorithms. For each algorithm,

   (a) report MLE $\widehat{\theta}$ and number of iteration,

   (b) make a table similar to that on Page 17 of the lecture notes "EM Algorithm". If the table is too long, you can just show the first and last few rows.

   (c) compare the two algorithms in terms of convergence rate.