Purpose: Explore advanced topics in optimization, regression, random number generation, and the bootstrap.

Assignment: Each problem involves mathematical and computational parts. For the computational parts, please turn in the simplest/shortest algorithm you can write that accomplishes the task. You may discuss the problems with other students, but all work must be your own. If you consult a book or online resource, cite it.

Problem 1: Weibull regression for a clinical trial

[Partly based on Lange "Numerical Analysis for Statisticians", Chapter 14, problem 23.]
A randomized trial comparing two treatments for ovarian cancer was conducted. It is described in

> Edmunson et al, "Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma vs. Minimal Residual Disease". *Cancer Treatment Reports*, 63:241-47, 1979.

The data are given in the `survival` package in `R`. Load this package and inspect the `ovarian` data frame. Type `?ovarian` to get a summary of the variable codings.

In survival analysis, we consider random variables $T \geq 0$ that represent survival times. Suppose T has density $f(t)$ and CDF $F(t)$. The *hazard function* is defined to be

$$h(t) = \lim_{s \to 0} \frac{\Pr(t < T \leq t + s \mid T > t)}{s} = \frac{f(t)}{1 - F(t)}.$$

This represents the instantaneous risk of death at each time t. The tail probability $1 - F(t)$ is known as the *survival function* and denoted $S(t) = 1 - F(t)$. In Cox's proportional hazards model, the hazard function has the form $h(t) = \lambda(t) \cdot e^{x'\alpha}$, where x is a $d \times 1$ vector of covariates and α is a vector of regression coefficients of corresponding dimension. The baseline hazard $\lambda(t)$ is a function of time t only and is common to all subjects.

This clinical trial involves right-censored survival times. This means that some patients are lost to the study before they die, and we only have the last time they were known to be alive. For example, if the survival time of subject i is censored at time t, then we only know that $T_i > t$. To formalize the notation for censoring, suppose we observe a time T_i and censoring indicator w_i for each subject i, and

$$T_i = \begin{cases} \text{time of death} & \text{if } w_i = 0 \\ \text{censoring time} & \text{if } w_i = 1. \end{cases}$$

In the Weibull proportional hazards model, $\lambda(t) = \beta t^{\beta-1}$, and so

$$S(t) = 1 - F(t) = e^{-t^\beta e^{x'\alpha}} \qquad \text{and} \qquad f(t) = \beta t^{\beta-1} e^{x'\alpha - t^\beta e^{x'\alpha}}.$$

a. Suppose you observe the possibly censored survival times $t_1, \ldots, t_m$, covariate vectors $x_1, \ldots, x_m$, and corresponding censoring indicators $w_1, \ldots, w_m$ for subjects $1, \ldots, m$. We wish to learn about the regression coefficients α and the Weibull parameter β. Write the likelihood and log-likelihood.

b. Write the gradient of the log-likelihood.

c. Show that the Hessian of the log-likelihood is

$$d^2\ell(\alpha, \beta) = -\left[\sum_{i=1}^m t_i^\beta e^{x_i'\alpha} \begin{pmatrix} x_i \\ \log t_i \end{pmatrix} \begin{pmatrix} x_i \\ \log t_i \end{pmatrix}' + (1 - w_i) \begin{pmatrix} 0 & 0 \\ 0 & \beta^{-2} \end{pmatrix}\right]$$

Derive an algorithm to perform Weibull proportional hazards regression.

d. Fit the data in the `ovarian` dataset using your algorithm. The covariate vector for each subject should consist of an intercept, age, residual disease, study arm, and ECOG performance. Find parameter estimates and their standard errors. Which treatment is better?

e. Extra Credit: Show that the log-likelihood is concave.

f. Extra Credit: Compare your results to the same regression model in `R`:

```
fit = survreg(Surv(futime, fustat) ~ age + resid.ds + rx + ecog.ps,
              data=ovarian, dist='weibull')
summary(fit)
```

Do you get the same estimates? The same standard errors? [Hint: see `?survreg` for the definition of the scale parameter.]

Problem 2: Poisson regression

Consider n independent observations $X_1, \ldots, X_n$ with

$$X_i \sim \text{Poisson}(\lambda_i) \quad \text{where} \quad \log \lambda_i = Z_i'\beta$$

is the subject-specific mean. Z_i and β are vectors of the same dimension and β is unknown.

a. Write the likelihood and log-likelihood of the data.

b. Derive a Newton-Raphson algorithm to estimate β by maximum likelihood.

c. Simulate observations as follows:

```
n = 100
d = 13
z = array(NA, dim=c(n,d))
for(i in 1:n) z[i,] = c(1, runif(d-1, -3,1))
beta = rnorm(d,mean=0,sd=1)
x = rpois(n, exp(z %*% beta))
```

Implement a routine for Poisson regression that does *not* use explicit matrix inversion, and instead uses the `solve` function with two arguments, as in `x = solve(A,b)` to solve the system $Ax = b$. Your routine should accept a vector of outcomes and a matrix of covariates as arguments: `poisreg(x,z)`. Your routine should report the asymptotic standard error for each estimated element of β. Check your results using `glm`.