Statistical Computing 1, Stat 590 | Homework 11 | Name: ______________________
Fall 2015 | Prof. Erik B. Erhardt

Part I. (120 points) Do all calculations in LaTeX + R + knitr. Insert computer text output and graphics to support what you are saying. For this assignment, all R code should well commented and be visible (`echo=TRUE`) in the document where you have written it.

(120pts) **1. Exponential regression**

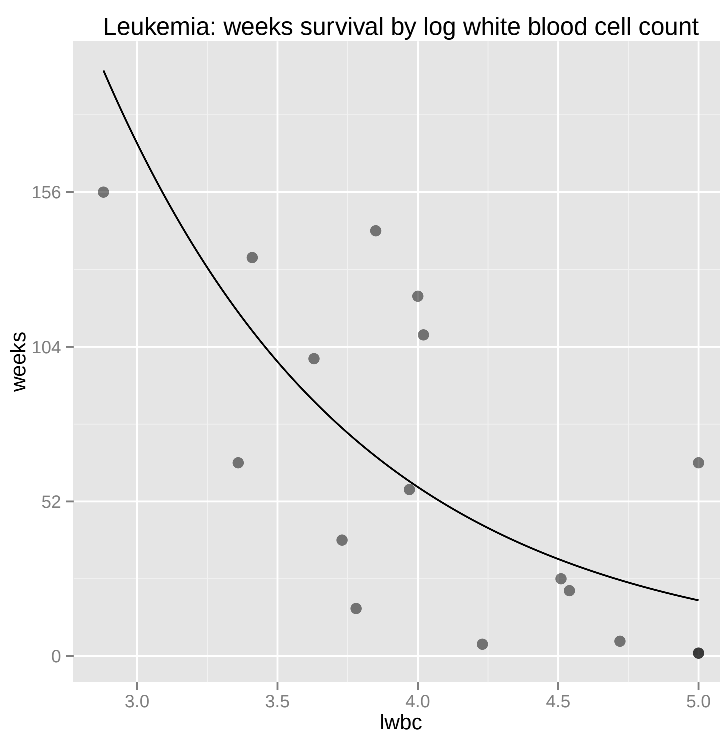
The data below from Feigl and Zelen (1965) are time to death, Y (weeks), in weeks from diagnois and $\log_{10}$(wbc) (the initial white blood cell count), x (lwbc), for 17 patients suffering from leukemia. The relation between Y and x is the main aspect of interest.

```
eleuk <- read.table("http://statacumen.com/teach/SC1/SC1_HW11_Exp-leuk.dat", header = TRUE)
```

	lwbc	weeks
1	3.36	65
2	2.88	156
3	3.63	100
4	3.41	134
5	3.78	16
6	4.02	108
7	4.00	121
8	4.23	4
9	3.73	39
10	3.85	143
11	3.97	56
12	4.51	26
13	4.54	22
14	5.00	1
15	5.00	1
16	4.72	5
17	5.00	65

Plot: Horizontal lines at 52-week (year) intervals.

```
library(ggplot2)
p <- ggplot(eleuk, aes(x = lwbc, y = weeks))
p <- p + scale_y_continuous(breaks = 52*c(0,1,2,3))
p <- p + stat_function(fun = f.exp.pred, args = list(out.FS$beta.MLE) )
p <- p + geom_point(alpha = 0.5, size = 3)
p <- p + labs(title = "Leukemia: weeks survival by log white blood cell count")
print(p)
```



A plot shows substantial random variation together with a tendency for Y to decrease with increasing x. Elaborate model fitting would be unnecessary for the analysis of these data in isolation. In particular, many different parametric representations of the systematic variations are consistent with the data.

We use the data to illustrate two main points. One is the use of general considerations to choose between alternative parametric regression relations. The other is the examination of the form of the random variation about a relation.

Three main aspects are involved in setting up a parametric description:

1. a specification of the form of systematic variation (mean), for example by giving the relation between the expected value $\mathrm{E}[Y_i]$ and x_i for the ith individual,
2. a specification of the general form of the random variation (residuals), and
3. a specification of the way in which systematic and random variation "combine", that is, by multiplication or by addition.

It is preferable for (1) to choose where possible a relation that gives positive values for $\mathrm{E}[Y_i]$ for all possible parameter values and values of x. From this point of view, the relation

$$\mathrm{E}[Y_i] \;=\; \beta_0 \exp\{\beta_1(x_i - \bar{x})\},$$

where $\bar{x} = \sum x_i/n$, is preferable to, for example, a linear relation

$$\mathrm{E}[Y_i] \;=\; \gamma_0 + \gamma_1(x_i - \bar{x}).$$

If the random contribution of the ith observation is ε_i, simplicity of interpretation and fitting, and inspection of the data, suggesting multiplicative combination; that is, we consider an interpretation in which the proportional variation around the mean has the same form for all x. It is then of interest to compare that distribution with the exponential distribution, partly because that is about the simplest very dispersed distribution for a positive quantity, partly because of the interpretation of the exponential distribution in terms of the properties of the completely random process, the Poisson process, partly because use of the exponential distribution much simplifies more detailed analysis, and consistency with exponential form is of great interest.
These considerations lead to the model

$$\mathrm{E}[Y_i] \;=\; \beta_0 \exp\{\beta_1(x_i - \bar{x})\}\varepsilon_i, \tag{1}$$

where ε_i is a random term of unit mean and, conceivably, exponentially distributed.
Suppose Y is a positive random variable with density

$$f(y|\mu) \;=\; \frac{1}{\mu}\exp\{-y/\mu\}, \quad y \geq 0,$$

then Y is an Exponential random variable with $\mathrm{E}[Y] = \mu$ and $\mathrm{Var}[Y_i] = \mu^2$. We write $Y \sim \text{Exponential}(\mu)$.
Note that Equation (1) is equivalent to $Y_i \sim \text{Exponential}(\mu_i)$, where

$$\begin{aligned} \mu_i \;&=\; \beta_0 \exp\{\beta_1(x_i - \bar{x})\}, \quad \text{or} \\ \log(\mu_i) \;&=\; \log(\beta_0) + \beta_1(x_i - \bar{x}). \end{aligned}$$

We will fit this model, but reparameterize the intercept so that

$$\log(\mu_i) \;=\; \beta_0 + \beta_1(x_i - \bar{x}), \tag{2}$$

that is, our intercept is the log of the original intercept.
Assume the data are a random sample $Y_1, Y_2, \ldots, Y_n$ independent with $Y_i \sim \text{Exponential}(\mu_i)$.

(a) (10 pts) Show that the log-likelihood function is

$$\ell(\underset{\sim}{\mu}|\underset{\sim}{y}) \quad = \quad -\sum_{i=1}^{n}\left(\frac{y_i}{\mu_i} + \log(\mu_i)\right).$$

For later reference, note that

$$\frac{\partial \ell}{\partial \beta_j} \quad = \quad \frac{y_i}{\mu_i^2} - \frac{1}{\mu_i}.$$

(b) (20 pts) In a standard exponential regression problem we assume that $\log(\mu_i)$ is linearly related to a vector $\underset{\sim}{x}_i^\top = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ of known covariates and an unknown vector $\underset{\sim}{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^\top$ of regression parameters via

$$\underset{\sim}{\mu}_i \quad \equiv \quad \underset{\sim}{\mu}_i(\underset{\sim}{\beta}) \quad = \quad \exp\{\underset{\sim}{x}_i^\top \underset{\sim}{\beta}\}$$

or

$$\log(\underset{\sim}{\mu}_i) \quad \equiv \quad \log(\underset{\sim}{\mu}_i(\underset{\sim}{\beta})) \quad = \quad \underset{\sim}{x}_i^\top \underset{\sim}{\beta} \quad = \quad \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

Typically, $x_{i1} = 1$ and if so β_1 is an “intercept effect”.
A goal here is to derive an NR/Fisher Scoring iteration procedure to compute the MLE of $\underset{\sim}{\beta}$. To this end, derive the derivatives of the mean vector and log-likelihood and derive the expected information:

1. first partial of mean vector

$$\frac{\partial \mu_i(\underset{\sim}{\beta})}{\partial \beta_j} \quad = \quad x_{ij}\mu_i(\underset{\sim}{\beta})$$

2. first partial of log-likelihood

$$\frac{\partial \ell(\mu(\underset{\sim}{\beta})|\underset{\sim}{y})}{\partial \beta_j} \quad = \quad \sum_{i=1}^{n} x_{ij}\left(\frac{y_i}{\mu_i(\underset{\sim}{\beta})} - 1\right)$$

3. second partial of log-likelihood

$$\frac{\partial^2 \ell(\mu(\underset{\sim}{\beta})|\underset{\sim}{y})}{\partial \beta_j \partial \beta_k} \quad = \quad -\sum_{i=1}^{n} x_{ij}x_{ik}\frac{y_i}{\mu_i(\underset{\sim}{\beta})}$$

4. Expected information

$$\mathbf{I} = \mathrm{E}\left(-\frac{\partial^2 \ell(\mu(\underset{\sim}{\beta})|\underset{\sim}{y})}{\partial \beta_j \partial \beta_k}\right) \quad = \quad \underset{\sim}{x}_j^\top \underset{\sim}{x}_k$$

(c) (15 pts) Using matrix notation, write the "structural" part of the model as

$$\log(\underset{\sim}{\mu}) \;=\; \mathbf{X}\underset{\sim}{\beta}$$

where $\mathbf{X}$ is an n-by-p design matrix with ith row $\underset{\sim}{x}_i^\top$ and $\log(\underset{\sim}{\mu})$ is an n-by-1 vector with ith element $\log(\mu_i)$.
Using obvious notation, derive the matrix representations of the derivatives of the mean vector and log-likelihood and the expected information, $\dot{\ell}(\underset{\sim}{\beta})$, $\ddot{\ell}(\underset{\sim}{\beta})$, $\mathbf{I}(\underset{\sim}{\beta}) = \mathrm{E}(-\ddot{\ell}(\underset{\sim}{\beta}))$.

1. first partial of log-likelihood

$$\dot{\ell}(\underset{\sim}{\beta}) \;=\; \left[\frac{\partial \ell}{\partial \beta_j}\right]_{p\times 1} \;=\; \mathbf{X}^\top \mathbf{D}^{-1}(\mu(\underset{\sim}{\beta}))(\underset{\sim}{y} - \mu(\underset{\sim}{\beta}))$$

   where $\mathbf{D}^{-1}(\mu(\underset{\sim}{\beta}))$ is an n-by-n diagonal matrix with diagonal elements μ_i, and $\underset{\sim}{y} = [y_1, y_2, \ldots, y_n]^\top$.

2. second partial of log-likelihood

$$-\ddot{\ell}(\underset{\sim}{\beta}) \;=\; \left[-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right]_{p\times p} \;=\; \mathbf{X}^\top \mathbf{D}(\underset{\sim}{y}/\underset{\sim}{\mu})\mathbf{X}$$

   where $\mathbf{D}(\underset{\sim}{y}/\underset{\sim}{\mu})$ is an n-by-n diagonal matrix with diagonal elements y_i/μ_i.

3. Expected information

$$\mathbf{I}(\underset{\sim}{\beta}) \;=\; \mathrm{E}(-\ddot{\ell}(\underset{\sim}{\beta})) \;=\; \mathbf{X}^\top\mathbf{X}.$$

   Recall that the MLE of $\underset{\sim}{\beta}$ solves the likelihood equations $\dot{\ell}(\underset{\sim}{\beta}) = \underset{\sim}{0}_p$, a p-by-1 vector of zeros, and that for large n

$$\hat{\underset{\sim}{\beta}} \;\dot{\sim}\; \mathrm{Normal}_p(\underset{\sim}{\beta}, \mathbf{I}^{-1}(\hat{\underset{\sim}{\beta}})),$$

   where the information matrix can be estimated by either $\mathbf{I}(\hat{\underset{\sim}{\beta}})$ or $-\ddot{\ell}(\underset{\sim}{\beta})$.

(d) (10 pts) Write down an iterative scheme to compute successive estimates of $\underset{\sim}{\beta}$ via Newton Rhapson and Fisher Scoring.

(e) (10 pts) Write a function to compute the log-likelihood function in Part (a) allowing for two vectors of input $\underset{\sim}{y}$ and $\underset{\sim}{\mu}$ and matrix $\underset{\sim}{X}$.

(f) (15 pts) Write a function to compute the MLE of $\underset{\sim}{\beta}$ for the exponential regression model $\log(\underset{\sim}{\mu}) = \mathbf{X}\underset{\sim}{\beta}$ using NR. Include a "line-search" or "step-size" adjustment as part of the routine. The "input" and "output" structure of this function should mimic my logistic regression functions in the notes.

(g) (10 pts) Incorporate a `method = "F"` option in your function in Part (f) to use either NR (`"N"`) or Fisher Scoring (`"F"`) and add the Fisher Scoring increment.

(h) (20 pts) Put it all together! Write a script that will read the survival data given at the start of the program, fit the model in Equation (2) as defined in the data description, and print and plot relevant summary output using both NR and Fisher Scoring.
Put some thought into the choice of appropriate starting values.
Qualitatively discuss the fit of the model, including an equation for the estimated mean survival time as a function of $\log_{10}(\text{wbc})$.

(i) (10 pts) Explore convergence of NR and Fisher Scoring as a function of the starting values. That is, does the choice of starting values affect whether the routine converges, and if so, how rapidly? Discuss.