FIRST YEAR EXAM

Spring 2000

Notes:

a. This is a closed book exam. No notes are permitted.

b. Please write your solutions on the paper provided. Put the problem number and your code on the top of each page that you turn in; write only on one side of the page. Start each problem on a new page.

c. It is to your advantage to show your work and explain your answers. Don't erase; draw a line through work you don't want graded.

d. All 6 problems will be graded.

e. You have 3 hours to finish.

1. Suppose the intensities of light from stars can be modeled as an exponential distribution with an unknown parameter $\lambda > 0$, with density:

$$f(x|\lambda) = \lambda e^{-\lambda x}, \text{ for } x > 0.$$

a. A device produces measurements of intensities, $X_1, X_2, \ldots, X_n$. Find the maximum likelihood estimator of the parameter λ.

b. Suppose the device can only detect intensities bigger than C where C is a known value, $C \geq 0$. Furthermore, only intensities for which X is greater than C are recorded. Find the density function of X given that $X > C$.

c. Suppose the prior distribution of λ is $Gamma(\alpha, \beta)$ with $\alpha, \beta > 0$, with density

$$\pi(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \text{ for } \lambda > 0.$$

Given observed data $X_1, \ldots, X_n$ from the distribution in **(b)**, find the posterior distribution of λ. *Note: if you could not do* **(b)**, *use the distribution from* **(a)**.

d. Using the posterior distribution from **c.**, what is the Bayes estimator of λ under squared error loss?

2. The following analysis addresses the question of sex bias in annual professional salaries, after taking into account other variables. There are $n = 44$ observations classified by three factors: `Sex`, a two-level factor (Sex=0 for Men, Sex=1 for Women); `Man`, a management level factor with two levels indicating the job level of the person (0=Lower, 1=Upper management); and `Ed`, a factor at three levels indicating the level of education of the person (0=High School, 1=Bac. Degree, 2=Graduate Degree). The other variables are `Salary`, the annual salary, and `Years`, the years in the job.

The following pages (4-5) give summaries for four different linear regression model fits, labeled `fit.1` to `fit.4`, where the response variable is the natural *log* of salaries. Three separate graphs of `Salary` against `Years` indicate the salaries classified according to each of the three factors (points are labeled by the level of the factor in each of the three graphs).

Notes: In the following analysis, factors are coded using dummy variables, so that adding a factor variable to a model simply changes the intercept by a parameter for each level of the factor relative to level 0. In each of these four models, the slope parameter on the predictor `Years` *does not vary with any of the factors. For example, in* `fit.2` *the model is estimated as* $9.5171 + 0.0334$`years` *for management level 0, and* $9.5171 + 0.2142 + 0.0334$`years` $= 9.7313 + 0.0334$`years` *for management level 1. Similarly, in* `fit.3` *the fitted model is* $9.4769 + 0.0339$`years` *for educational level 0, but* $9.4769 + 0.0762 + 0.0339$`years` $= 9.5531 + 0.0339$`years` *for educational level 1, and* $9.4769 + 0.0681 + 0.0339$`years` $= 9.545 + 0.0339$`years` *for educational level 2.*

a. Does the graph of `Salary vs Years in Job by Sex factor` on page 5 suggest likely salary differences between Men and Women? Briefly discuss why this graph alone may give a misleading indication of possible differentials.

b. Which of `fit.1` and `fit.2` gives a better predictive fit to the data, and why?

c. Why does `fit.3` have 40 degrees of freedom, whereas `fit.1` and `fit.2` have 41?

d. Does `fit.4` appear to be a useful model in explaining variations in salaries? Why or why not?

e. What does `fit.4` imply about differences in salary levels between men and women, taking into account the other factors? Do your conclusions seem reasonable in connection with the graph?

f. Remembering that the response is salary on the **log** scale, what does the estimated coefficient of 0.092 for `Ed1` imply for differences in salaries between people with Bac. degrees relative to those with just high school education?

g. What do you make of the fact that the estimated `Ed2` coefficient is significantly smaller than that for `Ed1`? Do the graphs give you any indication of why this perhaps surprising result arises?

SOME EXPLORATORY REGRESSION ANALYSES OF SALARY SURVEY DATA
FOR PROBLEM 2

```
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

 fit.1   <- lm( logsalary ~ Sex + years)
 summary(fit.1)

                Value Std. Error   t value  Pr(>|t|)
(Intercept)    9.5032    0.0639   148.6461    0.0000
        Sex   -0.0030    0.0465    -0.0650    0.9485
      years    0.0298    0.0075     3.9671    0.0003

Residual standard error: 0.243 on 41 degrees of freedom
Multiple R-Squared: 0.3053


>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

 fit.2   <- lm( logsalary ~ Man + years)
 summary(fit.2)

               Value Std. Error  t value Pr(>|t|)
(Intercept)   9.5171   0.0285   334.4820   0.0000
        Man   0.2142   0.0167    12.8125   0.0000
      years   0.0334   0.0032    10.5483   0.0000

Residual standard error: 0.1086 on 41 degrees of freedom
Multiple R-Squared: 0.8611


>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

 fit.3   <- lm( logsalary ~ Ed + years)
 summary(fit.3)

               Value Std. Error  t value Pr(>|t|)
(Intercept)   9.4769   0.0578   164.0321   0.0000
        Ed1   0.0762   0.0394     1.9356   0.0600
        Ed2   0.0681   0.0247     2.7531   0.0088
      years   0.0339   0.0065     5.2088   0.0000

Residual standard error: 0.2183 on 40 degrees of freedom
Multiple R-Squared: 0.4529
```

```
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

 fit.4   <- lm( logsalary ~ Sex+Man+Ed + years)
 summary(fit.4)

               Value Std. Error  t value Pr(>|t|)
(Intercept)   9.5051   0.0180   527.1676   0.0000
        Sex   0.0228   0.0131     1.7404   0.0899
        Man   0.2097   0.0108    19.4075   0.0000
        Ed1   0.0920   0.0123     7.4661   0.0000
        Ed2   0.0295   0.0080     3.7133   0.0007
      years   0.0362   0.0022    16.7819   0.0000

Residual standard error: 0.06769 on 38 degrees of freedom
Multiple R-Squared: 0.95

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
```

Figure 1: Graphs of Salary versus Years in Job by various factors for Problem 2

3. Suppose the distribution for observations $Y = (Y_1, \ldots, Y_n)'$ is multivariate normal,

$$Y|\beta, w \sim N_n(X\beta, w\, I_n)$$

where X is a known $n \times p$ full rank matrix, β is a vector of length p, and $w > 0$ is the unknown variance of Y_i. Also suppose that the prior distribution for β given X and w is multivariate normal with mean α (a p-dimensional vector) and variance-covariance matrix $w\tau(X'X)^{-1}$,

$$\beta|w \sim N_p(\alpha, w\tau(X'X)^{-1}),$$

where $\tau > 0$ and the hyperparameters τ and α are both known. The prior for w is given by

$$p(w) \propto 1/w, \text{ for } w > 0.$$

a. Derive the marginal posterior density of β (up to a normalizing constant).

b. Show that the posterior mean of β can be expressed as a weighted average of the least-squares estimate $\hat{\beta}$ and the prior mean α. *Note: You do not need to derive the posterior mean; just identify it from* **(a)**.

Hints:

$$X'Y = (X'X)\hat{\beta}, \qquad \text{where} \qquad \hat{\beta} = (X'X)^{-1}X'Y$$

$$\int_0^\infty w^{-c-1} \exp(-d/w)dw \;=\; \frac{\Gamma(c)}{d^c}, \text{ where } c > 0, d > 0.$$

4. A standard steroid test is used to randomly test athletes for steroid abuse. The test has a 1% "false positive" rate — so, for a randomly selected athlete undergoing the test, $\text{Prob}(x = 1|\theta = 0) = 0.01$ where: $x = 1$ indicates a positive test result, and $x = 0$ a negative test result; and $\theta = 1$ indicates a steroid user, $\theta = 0$ indicates a non-user.

a. An athlete tests positive. This leads one athletic official to state that

"Since the positive test result had only a 1% chance of occurring if she is a non-user, then there is about a 99% chance that she is a user."

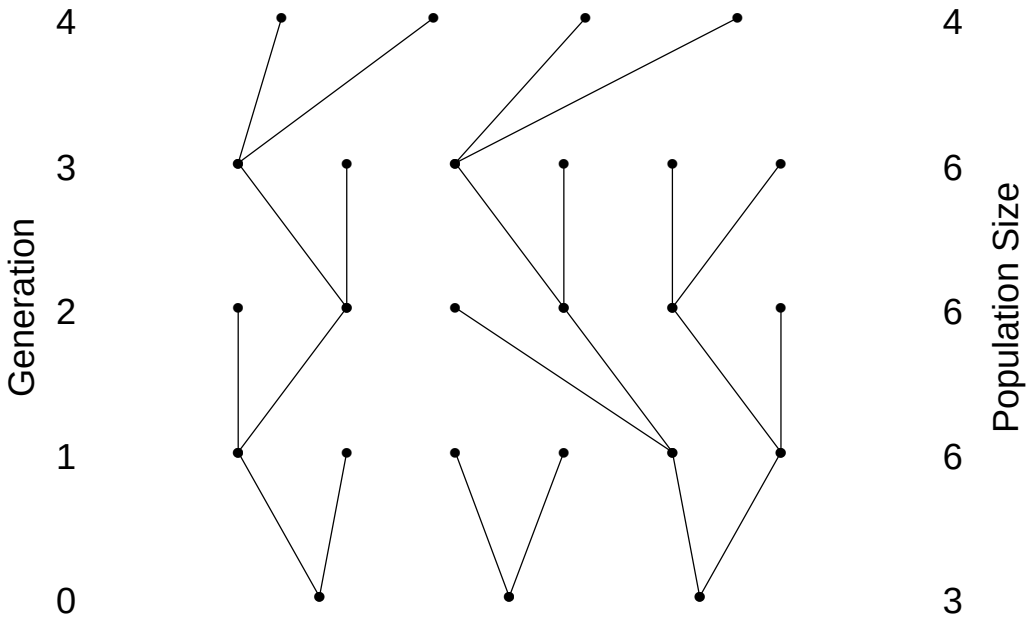
Explain why that reasoning is incorrect and an incomplete analysis of the implications of the test result.

b. A more enlightened official recognizes that we also need to know how accurate the test is among users; it is, in fact, perfectly accurate, so that $\text{Prob}(x = 1|\theta = 1) = 1$. This official then states that

"She is most probably a drug user because a positive result is 100 times more likely for a drug users than for non-users."

Explain why this reasoning is also incorrect and incomplete.

5. In a certain population, individuals may independently reproduce only one time, having either two offspring (with probability p) or no offspring (with probability $q = 1 - p$). Below is a figure which follows the family lines of a simple random sample of three individuals over four generations. The population at each generation is also given to the right of the figure. So, at generation 0, each individual had two offspring; at generation 1, 3 individuals had 0 offspring and 3 had 2 offspring; and so on. Whether or not the generation 4 individuals have offspring is not observed.



a. Given the above realization, give a sensible estimate for p, the chance an individual has exactly two offspring.

b. Is the estimate you gave above certain to converge to p as you continue to observe this population over successive generations? If so, explain why; if not, give a counterexample for which the estimate does not converge to p.

c. Conditional on having 3 individuals in the population at generation 0, what is the expected population size as a function of p for generation k, for any $k \geq 1$?

6. For a series of n days, airborne particles are measured by taking in a fixed volume of ambient air; the particles are deposited onto a filter and later weighed. Let X_i for $i = 1, \ldots, n$ denote the series of particle measurements and assume that these are independently and identically distributed from a distribution with cumulative distribution function $F(x) = 1 - (1/x)^{\beta}$, $x > 1$, and $\beta > 0$ and is 0 otherwise. The value 1 corresponds to the lower limit of detection for the problem.

a. What is the density of X?

b. Show that the joint distribution of the X's constitutes an exponential family. What is the natural parameter?

c. Find the MLE of β. What is the MLE of $1/\beta$?

d. Show that the MLE of $1/\beta$ is an unbiased estimator of $1/\beta$. Is the MLE of β an unbiased estimator of β?

e. Suppose that $g(X_1)$ is an unbiased estimator of β. How can the Rao-Blackwell Theorem be applied to construct an improved estimator of β? What properties would this new estimator have?