**STA 442/2101F 2009 Homework 2.** *due November 10 before 4 pm (revised date)*

**When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. DO NOT include in this summary printouts of computer code with the relevant selections highlighted.**

**All computer code used to obtain the results summarized in the response should be provided as an appendix. In this appendix you may highlight the relevant results.**

1. **STA 442/2101** Factorial analysis of Example J:
   In this question we will use the coded values $-1, 0, +1$ for $x_1$, $x_2$ and $x_3$, instead of taking logs of their numerical values.

   (a) You need to tell `R` that these variables are coded: the command for this is `factor`. The command `aov` will fit a linear model (similarly to `lm`), and it is often useful for balanced data such as we have here. Execute the command

   ```
   aov(log(cycles) ~ factor(x1) + factor(x2) + factor(x3))
   ```

   and summarize the results in an analysis of variance table. Does the analysis of variance table change if you enter the variables in a different order, e.g. `factor(x3)+factor(x2)+factor(x1)`?

   (b) Why are there 2 degrees of freedom for each of these factors in the analysis of variance table? What is the residual sum of squares and its degrees of freedom?

   (c) The sum of squares due to $x_1$ can be partitioned into a linear and quadratic component, using the contrasts $(-1, 0, +1)$ and $(+1, -2, +1)$. The sum of squares for a contrast vector $c$ is $SS_c = \{c_1(\bar{y}_{-1} - \bar{y}) + c_2(\bar{y}_0 - \bar{y}) + c_3(\bar{y}_{+1} - \bar{y})\}^2 r / \sum c_j^2$, where $\bar{y}_{-1}$ is the average of $\log(cycles)$ at the value $x_1 = -1$, and so on, and $r$ is the number of observations in $\bar{y}_{-1}$, i.e. 9. Carry out this partitioning for each of the three factors $x_1$, $x_2$ and $x_3$, and display the analysis of variance table with these entries.

   (d) What is the connection between the sums of squares in the table of (c) to those in Table J.3?

2. **STA 442** Randomized block design: Suppose we have $B$ blocks, each with $T$ experimental units, obtained in such a way that the units within blocks are expected to have similar responses under the same conditions. To each unit in a block we randomly assign one of $T$ treatments. The data in Table 1 show the weight gain on each of 4 diets (the treatments) for 5 litters (the blocks) of pigs.

Table 1: Weight gain in pigs, from Davison, 2003.

| | Litter, or block | | | | |
|---|---|---|---|---|---|
| Diet | 1 | 2 | 3 | 4 | 5 |
| I | 1.40 | 1.79 | 1.72 | 1.47 | 1.26 |
| II | 1.31 | 1.30 | 1.21 | 1.08 | 1.45 |
| III | 1.40 | 1.47 | 1.37 | 1.15 | 1.22 |
| IV | 1.96 | 1.77 | 1.62 | 1.76 | 1.88 |

(a) A standard linear model for this setting is

$$y_{tb} = \mu + \alpha_t + \beta_b + \epsilon_{tb}, \quad t = 1, \dots T, b = 1, \dots, B \tag{1}$$

with the assumption that $\epsilon_{tb}$ are independent, identically distributed random variables with mean 0 and variance $\sigma^2$. Show that this model can be expressed (with a slight abuse of notation) as

$$y = X\beta + \epsilon$$

where $y$ is an $n = TB \times 1$ vector, $X$ is an $TB \times T + B + 1$ matrix, and $\beta = (\mu, \alpha_1, \dots, \alpha_T, \beta_1, \dots, \beta_B)$ is the vector of unknown parameters in the mean. Give the form of $X$ explicitly for $T = 4$ and $B = 5$. What is the rank of $X$?

(b) Show that

$$\sum_{t=1}^{T}\sum_{b=1}^{B}(y_{tb} - \bar{y}_{..})^2 = B\sum_{t=1}^{T}(\bar{y}_{t.} - \bar{y}_{..})^2 + T\sum_{b=1}^{B}(\bar{y}_{.b} - \bar{y}_{..})^2 + \sum_{t=1}^{T}\sum_{b=1}^{B}(y_{tb} - \bar{y}_{t.} - \bar{y}_{b.} + \bar{y}_{..})^2. \tag{2}$$

(c) Construct the analysis of variance table based on (2), for the data in Table 1. The easiest way to do this in `R` is to use `aov` followed by `anova`.

(d) Is there evidence that the different diets lead to different weight gain? Explain.

3. **STA 442/2101** Logistic Regression: In fall quarter, 1973, there were 8,442 men who applied for admission to graduate school at the University of California, Berkeley, and 4,321 women. About 47% of the men and 35% of the women were admitted. On the basis of these numbers, an investigation was launched into the source of the apparent discrimination. Since admissions are made separately for each major, the admissions data were broken down by major. The data for the six largest majors, accounting for over one third of the total number of applicants, is given below.

|  | Men |  | Women |  |
|---|---|---|---|---|
|  | Number of | Percent | Number of | Percent |
| Major | applicants | admitted | applicants | admitted |
| A | 825 | 62 | 108 | 82 |
| B | 560 | 63 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |
| Total | 2691 | 44 | 1835 | 30 |

The data are available as UCBAdmissions in the DAAG package, in the form of $2 \times 2$ tables for each major. To put it into a form suitable for logistic regression, the following code works:

```
UCB = data.frame(admit = as.vector(UCBAdmissions[1, , ]),
      reject = as.vector(UCBAdmissions[2, , ]),
      Gender = rep(c("Male", "Female"),6),
     Dept = rep(LETTERS[1:6],rep(2,6)))
UCB$Gender = relevel(UCB$Gender, ref="Male")
UCB$total = UCB$admit + UCB$reject
UCB$phat = UCB$admit/UCB$total
```

The data frame UCB is now suitable for fitting logistic regression. The response variable phat is the observed proportion of admissions in each Gender/Dept category.

(a) Fit a logistic regression model $y_{ij} \sim \text{Binom}(n_{ij}, p_{ij})$, where $j = 1, ..., 6$ indexes departments and $i = 1, 2$ indexes gender.[1] Assume the logit link, and fit a logit-linear model with terms for department, for gender, and for their interaction, i.e.

$$\log\{p_{ij}/(1 - p_{ij})\} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

Give a summary table showing the analysis of deviance. Which factor(s) seem to influence the probability of admissions to Berkeley?

(b) Give the summary table of coefficient estimates, noting which constraints have been invoked. Choose 3 of the estimated coefficients as examples, and explain how they are to be interpreted.

(c) The overall summary table shows a substantial preference for males in the admissions (44% vs 30%). Based on the analysis in (a) and (b), do you find evidence of a gender bias in admissions? Explain.

4. **STA 442/2101** Suppose we have $n$ independent binary random variables $Y_1, \dots Y_n$, and

$$\Pr(Y_i = 1) = p_i(\beta) = \frac{\exp(x_i^T\beta)}{1 + \exp(x_i^T\beta)}, \quad \Pr(Y_i = 0) = 1 - p_i(\beta) = \frac{1}{1 + \exp(x_i^T\beta)} \qquad (3)$$

[1] There are several ways to fit logistic regression in R, check the details of the glm command carefully. Of course you are welcome to use other software if you prefer.

where $x_i$ is a $p \times 1$ vector of covariates and $\beta$ is a $p \times 1$ vector of unknown parameters.

(a) Construct the log-likelihood function for $\beta$ and show that it depends on $Y$ only through $S = \sum Y_i x_i$; i.e. S is a sufficient statistic for $\beta$.

(b) Show that the maximum likelihood estimate of $\beta$ based on a sample $(y_1, \dots y_n)$ is defined by the equation

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} p_i(\beta) x_i$$

where $p_i(\beta)$ is given in (3).

(c) Find the observed Fisher information function for $\beta$, defined as $j(\beta) = -\partial^2 \log L(\beta)/\partial\beta\partial\beta^T$.

5. **STA 2101 (comp)** Data has been collected on 2000 subjects who were recruited at age 20 for a study of whether or not smoking and alcohol consumption are related to the probability of early death. These people were followed for the next 30 years, after which the following binary (0/1) variables where available for each person (remarkably, the investigators where able to keep in touch with all 2000 subjects, so there is no missing data):

**smoked** 1 if the subject smoked at least 20 packs of cigarettes during the 30-year period of the study.

**drank** 1 if the subject consumed an average of at least one alcoholic drink per week during the 30-year study period.

**died** 1 if the subject died during the 30-year period of the study.

The investigators now wish to analyze the data, and draw conclusions. They believe that it is reasonable to assume that whether one subject died early is independent of whether any other subject died early, and you should assume that this is so in answering the questions below.

a) Is this an experiment or an observational study? Explain how the answer to this question affects what sort of conclusions the investigators might be able to draw.

b) The investigators fit a logistic regression model to look for a relationship between smoking and early death. If $p_i$ is the probability of early death for the $i$th subject, and $s_i$ is the variable indicating whether or not subject $i$ smoked, the logistic regression model for $p_i$ has the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 s_i$$

where $\beta_0$ and $\beta_1$ are model parameters that need to be estimated.

Is it possible that this model is wrong? If so, describe a situation in which it would be wrong, and explain how the investigators can check whether or not the model seems to be wrong.

c) The investigators also fit a logistic regression model to look for a relationship between both smoking and drinking and early death. In this model, if $p_i$ is the probability of early death for the $i$th subject, $s_i$ is the variable indicating whether or not subject $i$ smoked, and $a_i$ is the variable indicating whether or not subject $i$ drank alcohol, the logistic regression model for $p_i$ has the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 s_i + \gamma_2 a_i$$

where $\gamma_0$, $\gamma_1$, and $\gamma_2$ are model parameters that need to be estimated.

Is it possible that this model is wrong? If so, describe a situation in which it would be wrong, and explain how the investigators can check whether or not the model seems to be wrong.

d) Suppose that both the logistic regression models above seem to be appropriate. Is it possible for the estimate of $\beta_1$ in the model of (b) to be different from the estimate of $\gamma_1$ in the model of (c)? If so, how would you interpret such a difference?