Stat 175 problem for CWE August 2003

1. (Mixed effect model) Consider the model $y_{ij} = \mu + \alpha_i + e_{ij}$, where $i = 1, 2$, $j = 1, \ldots, n_i$, $n_1 = 2$, $n_2 = 3$, μ is fixed but unknown and all the other terms are independent with mean 0 and $\text{var}(\alpha_i) = \sigma_\alpha^2$, $\text{var}(e_{ij}) = \sigma_e^2$. Suppose we know $\sigma_\alpha^2/\sigma_e^2 = 2$.

   (a) Find the BLUE of μ as a linear combination of y_{ij}.

   (b) Find the BLUP of $\mu + \alpha_1$ as a linear combination of y_{ij}.

2. (Generalized linear model) The "Wisconsin Breast Cancer Database" was collected by Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison. A modified version is used in this problem. The samples consist of visually assessed nuclear features of fine needle aspirates (FNAs) taken from patients' breasts. Each sample has been assigned 9 attributes (attributes 2 to 10 below) by Dr. Wolberg. Each component of the attributes 2 to 8 is in the interval 1 to 10, with value 1 corresponding to a normal state and 10 to a most abnormal state. Attribute 9 has four levels (A, B, C and D), with A corresponding to normal state and D the most abnormal state. Attribute 10 has three levels (I, II and III), with I normal state and III most abnormal state. Attribute 1 designates whether the sample is benign or malignant. Malignancy is determined by taking a sample tissue from the patient's breast and performing a biopsy on it. A benign diagnosis is confirmed either by biopsy or by periodic examination, depending on the patient's choice.

   The attributes are as follows:

Field	Attribute
1	Class: benign or malignant
2	Clump Thickness
3	Uniformity of Cell Size
4	Uniformity of Cell Shape
5	Marginal Adhesion
6	Single Epithelial Cell Size
7	Bland Chromatin
8	Normal Nucleoli
9	Bare Nuclei
10	Mitoses

   There are 699 cases in the data and below are the first five:

```
  diagnosis thick size shape adhesion epi bland normal bare mitoses
1       Ben     5    1     1        1   2     3      1    A       I
2       Ben     5    4     4        5   7     3      2    D       I
3       Ben     3    1     1        1   2     3      1    A       I
4       Ben     6    8     8        1   3     3      7    B       I
5       Ben     4    1     1        3   2     3      1    A       I
```

A generalized linear model is used to determine μ, the probability that a sample is malignant. This model is fitted with the following commands in S-PLUS.

```
>options(contrasts = c("contr.treatment", "contr.treatment"))
>wbc.fit1 <- glm(diagnosis ~ thick + size + shape + adhesion + epi + bland +
        normal + bare + mitoses, family = binomial, data = wbc.dat)
>summary(wbc.fit1)$coef
                Value Std. Error  t value
(Intercept) -9.656270    1.17888 -8.19102
      thick  0.579212    0.14434  4.01282
       size -0.052947    0.20037 -0.26424
      shape  0.366902    0.22285  1.64639
   adhesion  0.221204    0.11197  1.97564
        epi  0.131177    0.15540  0.84414
      bland  0.501394    0.16738  2.99549
     normal  0.115204    0.11005  1.04680
      bareB  2.819494    0.73669  3.82727
      bareC  2.408169    0.84291  2.85696
      bareD  3.455476    0.72878  4.74142
  mitosesII  0.656286    0.83318  0.78769
 mitosesIII  4.130099    4.16433  0.99178

> summary(wbc.fit1)$cor[9:11,9:11]
        bareB   bareC   bareD
bareB 1.00000 0.32585 0.32503
bareC 0.32585 1.00000 0.28336
bareD 0.32503 0.28336 1.00000
```

(a) Give the link function and deviance function that is used in this model. What is the degree of freedom of the residual deviance?

(b) Interpret the coefficient of bareB. Construct an approximate 95% confidence interval for the odds ratio (bareA vs bareB) of having a benign sample.

(c) The odds ratio (bareA vs bareB) of having a benign sample could be computed separately for each of the three subgroups of patients with mitoses level at I, II and III respectively. Assuming the linear logistic model fitted is correct, which subgroup of patients has the largest odds ratio?

(d) If we fitted the following model instead, what is the degree of freedom of the residual deviance?

```
>wbc.fit2 <- glm(diagnosis ~ thick + size + shape + adhesion + epi + bland +
        normal + bare* mitoses, family = binomial, data = wbc.dat)
```

(e) Construct 95% simultaneous confidence intervals for all pairwise contrasts of Bare Nuclei using Bonferroni method. You only need to give explicit results for the contrast between bareB and bareC. (You can use 2.646 as the t-value, but you need to give the α and degrees of freedom of t-distribution to get full credit.)

(f) A patient has the following attributes: (thick: 2, size: 3, shape: 1, adhesion: 4, epi: 2, bland: 5, normal: 7, bare: C, mitoses: II). Estimate the probability that her sample is benign.

(g) One purpose of this study is to find a way to classify patients into benign and malignant groups based on the nine attributes. Below is a table of the predicted probability from the model and the actual status of the sample:

```
> table(cut(fitted(wbc.fit1),pretty(fitted(wbc.fit1), 10)),wbc$diag)
               Ben Mal
0.0+ thru 0.1 435   1
0.1+ thru 0.2   8   3
0.2+ thru 0.3   1   1
0.3+ thru 0.4   2   0
0.4+ thru 0.5   0   2
0.5+ thru 0.6   2   4
0.6+ thru 0.7   1   7
0.7+ thru 0.8   1   5
0.8+ thru 0.9   4  14
0.9+ thru 1.0   4 204
```

Use the above information to construct a rule to classify the patients such that it minimize the total misclassification rate for the sample, and compute the misclassification rate. Based on your rule, is the sample from the patient in (f) benign or malignant?