Biostatistics 513 Spring 2001

Final Exam NAME:

INSTRUCTIONS: Please answer the questions in the space provided. If you are unclear about any question then please contact me for clarification. The exam will end at 10:20am. *This exam is to be completed without use of notes or text.*

There are a total of 25 questions (as parts to 3 main questions) each worth 4 points. Partial credit will be given. Calculation details need not be completed to obtain full credit (i.e. $\exp(-0.257 + 1.034)$ is a sufficient answer).

1. Choose one response for each of the following questions:

1(a) [4pts] In a logistic regression model, $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X1 + \beta_2 X2$, where $X1 = 0$ or 1 and $X2 = 0$ or 1, the disease odds ratio comparing $(X1 = 1, X2 = 1)$ to $(X1 = 0, X2 = 0)$ is given by $\exp(\ \beta_1\ ) \cdot \exp(\ \beta_2\ )$.

◯ True
◯ False

1(b) [4pts] If we have prospective data (follow-up data), then we can use logistic regression to estimate both the probability of disease and relative risks.

◯ True
◯ False

1(c) [4pts] When using Kaplan-Meier to estimate a survival curve, the censored observations are not used since the exact survival time is unknown, and therefore these observations can be excluded from the data.

◯ True
◯ False

1(d) [4pts] The hazard function, $h(t)$, represents the probability of dying before time t.

◯ True
◯ False

2. Accurate data concerning the prevalence and severity of retinopathy and associated risk factors are of importance in planning a well-coordinated approach to the public health problem posed by this complication of diabetes. Diabetic retinopathy is one of the leading causes of blindness in people aged 20 to 74 years in the United States. The disease is characterized by the appearance of small hemorrhages in the retina followed by detrimental progression. Identifying patients that may be at high risk of moderate or severe retinopathy is important for advising in opthalmologic care. Such data are also helpful in planning future studies, such as controlled clinical trials of treatments for diabetes and for diabetic retinopathy. Data on $n = 720$ diabetics whose age-at-onset was less than 30 years of age was collected and analyzed in Klein et al. (1984) *Archives of Opthalmology* and is discussed below. The measured variables include:

DISEASE = 1 if either the left or the right eye has moderate or severe retinopathy, 0 otherwise
AGE(1) = 1 if age-at-diagnosis for diabetes is < 10 years, 0 otherwise
AGE(2) = 1 if age-at-diagnosis for diabetes is $10 - 19$ years, 0 otherwise
AGE(3) = 1 if age-at-diagnosis for diabetes is ≥ 20 years, 0 otherwise

DUR(1) = 1 if diabetes duration < 10 years, 0 otherwise
DUR(2) = 1 if diabetes duration ≥ 10 years, 0 otherwise

GHEMO(1) = 1 if glycosylated hemoglobin $< 12\%$, 0 otherwise
GHEMO(2) = 1 if glycosylated hemoglobin $\geq 12\%$, 0 otherwise

Glycosylated hemoglobin measures the average glucose control for the previous two to three months and provides a measure of normal or abnormal control as a percentage.

The following scientific questions were posed by the investigators:

1. Is the age-at-diagnosis predictive of more severe disease at a fixed follow-up (duration)? That is, are subjects with earlier onset (diagnosis) more likely to develop serious retinopathy?

2. Is glycosylated hemoglobin level associated with disease, and does this association depend on the duration of diabetes?

The following shows logistic regression models that were fit to the study data using DISEASE as the response variable (ie. Y):

Model 0

```
Log likelihood = -437.59
------------------------------------------------------------------------------
         |  Coef.   Std. Err.   z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  AGE(2) |  0.158    .197      0.802   0.422      -0.228    0.544
  AGE(3) |  0.359    .216      1.660   0.097      -0.064    0.782
   _cons | -1.021    .153     -6.667   0.000      -1.321   -0.720
------------------------------------------------------------------------------
```

Model 1

```
Log likelihood = -326.84
------------------------------------------------------------------------------
         |  Coef.   Std. Err.    z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  AGE(2) |  0.415    .227      1.823   0.068      -0.031   0.861
  AGE(3) |  0.691    .255      2.702   0.007       0.189   1.192
  DUR(2) |  3.017    .260     11.582   0.000       2.506   3.527
GHEMO(2) |  0.178    .194      0.916   0.360      -0.202   0.558
   _cons | -3.357    .315    -10.645   0.000      -3.975  -2.738
------------------------------------------------------------------------------
```

Model 2

```
Log likelihood = -326.70
------------------------------------------------------------------------------
                |  Coef. Std. Err.    z     P>|z|    [95% Conf. Interval]
----------------+-------------------------------------------------------------
         AGE(2) |  0.408   .228      1.792   0.073      -0.038    0.855
         AGE(3) |  0.695   .255      2.718   0.007       0.193    1.197
         DUR(2) |  3.177   .421      7.544   0.000       2.351    4.002
       GHEMO(2) |  0.401   .490      0.819   0.413      -0.559    1.362
DUR(2)*GHEMO(2) | -0.267   .535     -0.499   0.618      -1.316    0.781
          _cons | -3.490   .423     -8.233   0.000      -4.321   -2.659
------------------------------------------------------------------------------
```

2(a) [4pts] Using Model 0 give an estimate of the crude odds ratio that compares disease risk for AGE(3) versus AGE(2).

2(b) [4pts] Use Model 1 to give an estimate of the adjusted odds ratio comparing AGE(2) to AGE(1) that controls for both duration of diabetes and the level of glycosylated hemoglobin.

2(c) [4pts] In order to answer the first scientific question (i.e. age-at-diagnosis as a risk factor) does diabetes duration and/or glycosylated hemoglobin appear to be important confounders? Justify your response.

2(d) [4pts] Use Model 1 to give an estimate of the probability of disease for a subject with age-at-diagnosis of 11, who has had diabetes for 12 years, and who has a glycosylated hemoglobin of 10.3%.

2(e) [4pts] In Model 2, give an interpretation of the coefficient of GHEMO(2) (Note: the estimated coefficient in this model is 0.401).

2(f) [4pts] Using Model 2 what is the estimated odds ratio comparing GHEMO(2)=1 to GHEMO(2)=0 for subjects with more than 10 years of diabetes duration (ie. DUR(2)=1) who are less than 10 years old at the time of diagnosis (ie. AGE(2)=AGE(3)=0)?

2(g) [4pts] Using these logistic regression models answer the second research question: "is glycosylated hemoglobin associated with disease and does this association depend on the duration of diabetes." Justify your response.

Further analysis considered use of logistic regression with the variables of interest modelled in their original measured scale:

AGEDX = age-at-diagnosis in years
DUR = duration of diabetes in years
GHEMO = glycosylated hemoglobin in percent

The following model was fit:

Model 2

```
Log likelihood = -361.57
------------------------------------------------------------------------------
         |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
   AGEDX |      0.0176   .0120871     1.458   0.145     -.0060733    .0413071
     DUR |      0.1152   .0109081    10.563   0.000       .093847    .1366058
   GHEMO |      0.0946   .0353849     2.676   0.007       .0253318   .1640381
   _cons |     -3.9283   .5500088    -7.142   0.000     -5.006347   -2.850352
------------------------------------------------------------------------------
```

2(h) [4pts] Give an interpretation of the estimated coefficient for diabetes duration (DUR) in Model 2. (Note: the coefficient is estimated as 0.1152).

2(i) [4pts] What are the assumptions that are being made in Model 2 and how would you check these assumptions?

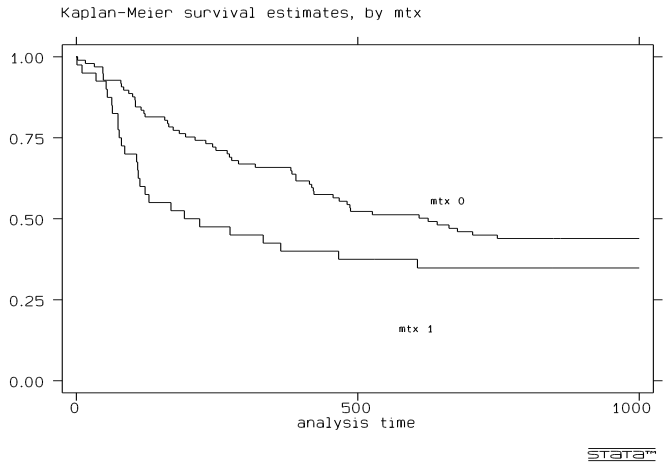
2(j) [4pts] Your colleague suggests that you would have more power in your logistic regression analysis if instead of using a single DISEASE response, defined as whether either the left or the right eye had moderate or severe disease, you used a pair of responses for each subjects using one response for each eye. Thus, with this suggestion each subject would contribute two observations - one an indicator of whether the left eye had moderate/severe disease and one an indicate of whether the right eye had moderate/severe disease. Comment on whether you agree that this is a wise suggestion. Justify your response.

3. Bone marrow transplants are a standard treatment for acute leukemia. A study was conducted with $n = 137$ subjects to examine the effects of the drugs Busulfan (BU) and cyclophosphamide (CY) as treatments in preparation for transplantation (Copelan et al. 1991, *Blood*). All patients received BU and CY. Interest in additional factors that may predict patient outcomes motivated an analysis of the survival time after transplantation and its association with other measured factors including:

MTX = 1 if the patient is treated with methotrexate,
0 otherwise.

The survival time of interest for the following analysis is the "disease-free time" defined as the length of time (in days) from transplantation to either death or disease relapse.

3(a) [4pts] The following plot shows the Kaplan-Meier curves for the two groups of patients defined by their MTX treatment status:



Based on these curves, estimate the percent that die or relapse by 250 days for each group.

death/relapse by day=250 for MTX=1 $\approx$

death/relapse by day=250 for MTX=0 $\approx$

3(b) [4pts] Describe how would you test whether the 250 day survival probabilites are different for these two groups (ie. MTX=1 versus MTX=0).

3(c) [4pts] A log-rank test was performed on these data and the value of the chi-square statistic was 3.40 (with df=1) and p-value 0.063. State the null and alternative hypotheses for this test:

H_0:

H_1:

3(d) [4pts] A weighted log-rank test was also performed using Wilcoxon weighting (ie. weighting proportional to the risk-set size – called the Peto test by Kleinbaum, and the Wilcoxon-Breslow by STATA) and a p-value of 0.0161 was obtained (chi-square statistic of 5.79). Explain why this result differs from the log-rank test and refer to the Kaplan-Meier curves to justify.

Regression analysis was used to explore the risk/benefit associated with MTX in addition to the prognostic value of other measured variables. The additional covariates include:

CMV = 1 if infected with cytomegalovirus, 0 otherwise
GROUP(1) = 1 disease group is lymphoblastic leukemia, 0 otherwise
GROUP(2) = 1 disease group is low risk myelotic leukemia, 0 otherwise
GROUP(3) = 1 disease group is high risk myelotic leukemia, 0 otherwise
AGE30 = (age - 30) in years
WAIT1 = (wait - 1) in years, where wait is the time from diagnosis to transplantation
MALE = 1 if subject is male and 0 if female

Consider the following Cox regression models:

Model 0

```
Log likelihood  =   -353.20
-------------------------------------------------
         |   Coef.   Std. Err.      z     P>|z|         p(PH)
---------+---------------------------------------
     MTX |    0.355    0.258      1.375   0.169         0.0235
     CMV |   -0.029    0.245     -0.121   0.904         0.7415
GROUP(2) |   -0.751    0.324     -2.312   0.021         0.9282
GROUP(3) |    0.377    0.306      1.231   0.218         0.3496
   AGE30 |   -0.00003  0.013     -0.002   0.998         0.5779
    MALE |   -0.195    0.235     -0.830   0.407         0.2954
   WAIT1 |   -0.132    0.141     -0.938   0.348         0.3735
-------------------------------------------------
```

Note: the column labeled `p(PH)` is the test of the proportional hazards assumption based on the Schoenfeld residuals.

Model 1

```
Log likelihood  =   -352.47
-------------------------------------------
         |  Coef.   Std. Err.   z     P>|z|
---------+---------------------------------
     MTX |    0.394   0.269      1.461   0.144
     CMV |   -0.016   0.247     -0.067   0.947
GROUP(2) |   -0.765   0.326     -2.344   0.019
GROUP(3) |    0.374   0.317      1.181   0.238
   AGE30 |   -0.002   0.013     -0.171   0.864
    MALE |   -0.244   0.242     -1.008   0.314
   WAIT1 |   -0.267   0.232     -1.150   0.250
 AGE30^2 |    0.001   0.000      1.025   0.305
 WAIT1^2 |    0.046   0.062      0.743   0.458
-------------------------------------------
```

Note: the variable

```
AGE30^2 = AGE30*AGE30
```

is the quadratic term for AGE30, and the variable

```
 WAIT1^2 = WAIT1*WAIT1
```

is the quadratic term for WAIT1.

3(e) [4pts] Write the Cox regression model that corresponds to Model 0. Define the necessary components of this model.

3(f) [4pts] What are the two key assumptions that are made when using Cox regression?

(1):

(2):

3(g) [4pts] Use Model 0 to estimate a hazard ratio comparing subjects with MTX=1 and CMV=1 to subjects with MTX=0 and CMV=0, who are otherwise equivalent in terms of the other covariates (GROUP, AGE30, MALE, WAIT1).

3(h) [4pts] Model 1 includes quadratic terms for the continuous covariates AGE30 and WAIT1. Is there evidence that these variables are not adequately modelled with just the linear terms? Justify your response based on appropriate statistics.

Since there is evidence that MTX violates the proportional hazards assumption, modified regression models were fit stratifying on MTX:

Model 2

```
Log likelihood  =   -304.11                    Stratified on MTX
------------------------------------------------------
         |   Coef.   Std. Err.      z    P>|z|
---------+--------------------------------------------
     CMV |  -0.067    0.245     -0.275   0.783
GROUP(2) |  -0.743    0.327     -2.271   0.023
GROUP(3) |   0.398    0.308      1.294   0.196
   AGE30 |  -0.001    0.013     -0.119   0.905
    MALE |  -0.248    0.235     -1.054   0.292
   WAIT1 |  -0.129    0.141     -0.917   0.359
------------------------------------------------------
```

Model 3

```
Log likelihood  =   -304.10                    Stratified on MTX
------------------------------------------------------
          |  Coef.   Std. Err.      z    P>|z|
----------+-------------------------------------------
      CMV |  -0.059    0.255     -0.232   0.817
 GROUP(2) |  -0.740    0.328     -2.255   0.024
 GROUP(3) |   0.398    0.308      1.292   0.196
    AGE30 |  -0.003    0.021     -0.174   0.862
     MALE |  -0.250    0.236     -1.061   0.289
    WAIT1 |  -0.131    0.142     -0.923   0.356
CMV*AGE30 |   0.003    0.027      0.127   0.899
------------------------------------------------------
```

3(i) [4pts] Give an interpretation of the coefficient of CMV in Model 2.

3(j) [4pts] Using Cox regression and Model 2, if the 1-year survival for (MTX=1, CMV=0, GROUP=1, AGE30=0, MALE=0, WAIT1=0) is estimated as 0.63 (i.e. this is the baseline 1-year survival defined by the reference category for each variable, $\widehat{S}_0(1yr) = 0.63$ for the MTX=1 strata), then based on Model 2, what would be the estimate of the 1-year survival probability for a person that was otherwise equivalent except they were GROUP=2 – that is they are MTX=1 and $X_j = 0$ for all other covariates except GROUP?

3(k) [4pts] Using Model 3 give an estimate of the hazard ratio that compares death/relapse risk comparing a CMV infected (CMV=1), Male, Age=40 to an uninfected (CMV=0), Male, Age=40 where both are (MTX=1, GROUP=2, WAIT1=0).