

**ADVANCED STATISTICAL METHODOLOGY (STAT 526)**  
**FINAL EXAM (REC 108)**  
**8:00-10:00AM, THURSDAY, MAY 03, 2007**

There are totally 43 points in the exam. The students with score higher than or equal to 40 points will receive 40 points. Please write down your name and student ID number below.

**NAME:** \_\_\_\_\_  
**ID:** \_\_\_\_\_

1. (10 points). The data describe the relationship between mental impairment and parent's socioeconomic status for a sample of residents of Manhattan in 1978. It recorded the count of residents according to their parents socioeconomic status (from "A"(high) to "F"(low)) and their mental status (from "Well" to "Impaired"). The R output is given.
  - (a) (2 points). Do you think the Mental Status is independent of the parent socioeconomic status. Why?
  - (b) (2 points). A linear-by-linear association term (ll) was fitted in the model. Does this model fit the data. Why?
  - (c) (2 points). A multinomial model was fitted and the output is given. Calculate the estimate probabilities when the Parent status is "A".
  - (d) (2 points). A proportional odds model was fitted. Does this model fit the data (hint: the multinomial model is the saturated model).
  - (e) (2 points). What are the estimated probabilities for Parent status "A" in the proportional odds model?

2. (12 points). The 50 samples of some kind of small trees were chosen and measured the weight. Those samples was selected from 5 different fields with 10 from each. It is known that the sample mean is  $\bar{y}_{..} = 4.6449$ , and the groups samples means are  $\bar{y}_1 = 3.0086$ ,  $\bar{y}_2 = 1.4064$ ,  $\bar{y}_3 = 4.2793$ ,  $\bar{y}_4 = 2.4246$  and  $\bar{y}_5 = 12.1059$  respectively.

(a) (3 points) Complete the following table.

	Df	SS	MS	F
Filed	4	?	?	?
Residuals	45	379.65	8.44	
Total	49	?		

(b) (2 points) Suppose one fits the data into a one-way random effect model. Specify this model and display the estimation of the parameters.

(c) (2 points) What could happen for the  $F$ -test of the significance of the random effect.

(d) (2 points) Estimate parameters if one assumes  $y_{ij}$  are iid  $N(\mu, \sigma^2)$ .

(e) (3 points) Specify the important steps for the  $p$ -value of a loglikelihood ratio test statistic in a bootstrap method of testing  $H_0 : \sigma_\alpha^2 = 0$ .

3. (13 points). The data reports the survival study of 40 males with late stage larynx cancer (stage 3 and stage 4). The three variables are STAGE of larynx cancer, TIME to death in months (30 days) and CENSOR indicator (1=death and 0=dropoff). The data for stage 3 are 0.3, 0.3, 0.5, 0.7, 0.8, 1.0, 1.3, 1.6, 1.8, 1.9, 1.9, 3.2, 3.5, 3.7+, 4.5+, 4.8+, 4.8+, 5.0, 5.0+, 5.1+, 6.3, 6.4, 6.5+, 7.8, 8.0+, 9.3+ and 10.1+. The data for stage 4 are 0.1, 0.3, 0.4, 0.8, 0.8, 1.0, 1.5, 2.0, 2.3, 2.9+, 3.6, 3.8, 4.3+. The R output is given behind.

(a) (3 points). Compute the Kaplan-Maier estimate of the survival function for stage 4.

(b) (2 points). Is  $\hat{S}(10) = 0$  for either stage? Why.

(c) (2 points). What is the null hypothesis of the logrank test. Does the logrank test claim the two survival functions are significantly different.

(d) (2 points). Suppose the survival time follows an exponential distribution. Estimate the expected survival time for both stages.

(e) (2 points). Suppose the survival time follows a Weibull distribution. What are the estimate of hazard functions for both stages.

(f) (2 points). Under the assumption that the survival time follows a Weibull distribution, do you accept that the survival time also follows an exponential distribution.

4. (8 points) The data reports Longitude and Latitude coordinates at which 37 the Class of 37 storms reached hurricane strength for two classifications of hurricanes—Baro hurricanes and Trop hurricanes. The output for the classification tree method is given.
- (a) (2 points) Explain the output and state the total numbers of the observations of the two classes in this dataset.
- (b) (2 points) State the number of observations as well as the the numbers of two classes on the left leaf in Figure 1.
- (c) (2 points). Compute the misclassification rate. What is the deviance of the tree.
- (d) (2 points). A logistic linear regression method which takes longitude and latitude as linear predators was fitted for the data and the fit was bad. Refer to the tree method and explain why.

Problem 1.

```
> g <- glm(Count~Parent+Mental,fam=poisson,data=mental)
> g$dev
[1] 47.41785
> g$df.residual
[1] 15
> g1 <- glm(Count~Parent+Mental+ll,fam=poisson,data=mental)
> g1$dev
[1] 9.895124
> g1$df.residual
[1] 14
> 1-pchisq(47.41,15)
[1] 3.164579e-05
> 1-pchisq(9.89,14)
[1] 0.7701807
> library(nnet)
> mmod
Call:
multinom(formula = y ~ as.factor(Parent))
```

Coefficients:

	(Intercept)	B	C	D	E	F
2	0.3844	0.1158	0.2265	0.2877	0.60676	0.8337
3	-0.0984429	0.0444	0.2298	0.1656	0.50389	1.0429
4	-0.3302366	-0.0239	0.3815	0.5969	1.10341	1.5484

```
> pmod <- polr(Mental~Parent, weights=Count,data=mental)
> summary(pmod)
```

Re-fitting to get Hessian

Call:

```
polr(formula = Mental ~ Parent, data = mental, weights = Count)
```

Coefficients:

	Value	Std. Error	t value
ParentB	-0.01697161	0.1607754	-0.1055610
ParentC	0.20817784	0.1548058	1.3447680
ParentD	0.29901019	0.1458232	2.0504985
ParentE	0.56676240	0.1583881	3.5783135
ParentF	0.82384896	0.1662160	4.9564958

Intercepts:

	Value	Std. Error	t value
Well Mild	-1.2039	0.1193	-10.0917
Mild Moderate	0.4953	0.1150	4.3065

Moderate|Impaired 1.5041 0.1203 12.5064

Residual Deviance: 4449.382

AIC: 4465.382

> c(mmod\$dev,mmod\$edf)

[1] 4441.554 18.000

> c(pmod\$dev,pmod\$edf)

[1] 4449.382 8.000

Problem 3.

Call:

survdiff(formula = Surv(Time, Censor) ~ Stage, data = larynx34)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Stage=3	27	17	21.8	1.06	5.2
Stage=4	13	11	6.2	3.71	5.2

Chisq= 5.2 on 1 degrees of freedom, p= 0.0226

Call:

survreg(formula = Surv(Time, Censor) ~ factor(Stage), data = larynx34,  
dist = "exponential")

	Value	Std. Error	z	p
(Intercept)	1.83	0.243	7.55	4.35e-14
factor(Stage)4	-1.06	0.387	-2.74	6.19e-03

Scale fixed at 1

Exponential distribution

Loglik(model)= -67.6 Loglik(intercept only)= -71

Chisq= 6.7 on 1 degrees of freedom, p= 0.0097

Number of Newton-Raphson Iterations: 4

Call:

survreg(formula = Surv(Time, Censor) ~ factor(Stage), data = larynx34,  
dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	1.8392	0.255	7.216	5.35e-13
factor(Stage)4	-1.0728	0.407	-2.638	8.34e-03
Log(scale)	0.0409	0.161	0.253	8.00e-01

Weibull distribution

Loglik(model)= -67.6 Loglik(intercept only)= -70.6

Chisq= 5.96 on 1 degrees of freedom, p= 0.015

Problem 4.

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 37 18 BARO (0.5135135 0.4864865)

2) Longitude>=67.75 10 0 BARO (1.0000000 0.0000000) \*

3) Longitude< 67.75 27 9 TROP (0.3333333 0.6666667)

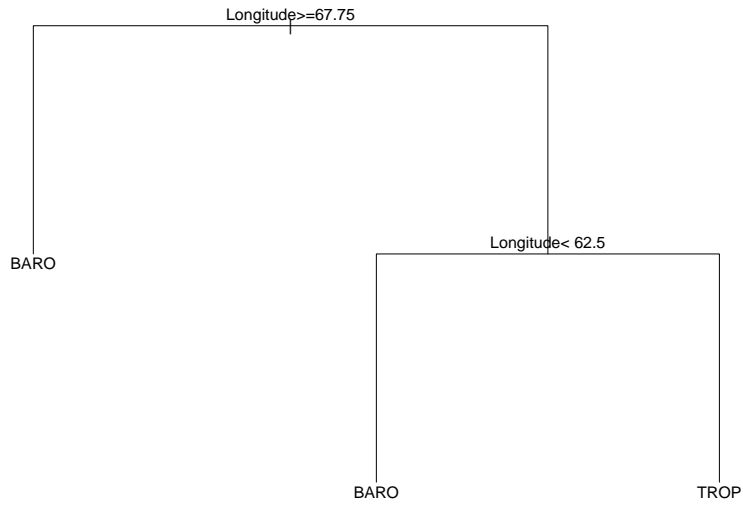


Figure 1: Tree in Problem 4

- 6) Longitude < 62.5 9 0 BARO (1.0000000 0.0000000) \*
- 7) Longitude >= 62.5 18 0 TROP (0.0000000 1.0000000) \*