2020

Math 485 Final Fall 2010 Name_____

YOU ARE BEING GRADED FOR WORK, NOT JUST THE FINAL ANSWER.
Eight sheets of notes.

1) Consider a $2 \times 2 \times k$ table with variables X, Y and Z. The study compared gender (boy and girl) (X), on whether the person passed a test (yes or no) (Y). Results from whether the person has low (less than 8 hours) or high (more than 8 hours) amount of sleep (Z). Suppose that CMH = 3.7013 and BD = 0.1501.

a) Perform the 4 step CMH test for conditional independence of X and Y given Z.

i) H_0 $\theta_{XY(1)} = \theta_{XY(2)} = 1$ H_A not H_0

ii) CMH = 3.7013

iii) $\begin{array}{c|cc} df & .1 & .05 \\ \hline 1 & 2.71 & 3.84 \end{array}$ $.05 < pval < .1$

iv) fail to reject H_0 X = gender and Y = passed test are conditionally independent given Z = amount of sleep.

b) Perform the 4 step Breslow Day test for homogeneity.

i) $H_0 \theta_{XY(1)} = \theta_{XY(2)}$ H_A not H_0 $(\theta_{XY(1)} \neq \theta_{XY(2)})$

ii) BD = 0.1501

iii) $df = k-1 = 1$ $\begin{array}{c|cc} & .25 & \\ \hline 1 & 1.32 & \end{array}$ $pval > .25$

iv) fail to reject H_0 there is homogeneous X-Y (gender - passed test) association given Z = amount of sleep

29

	B1	B2	B3	B4
df	945	965	968	976
# of predictors	55	35	32	24
# with $0.01 \leq$ Wald p-value ≤ 0.05	5	4	3	0
# with Wald p-value > 0.05	8	0	0	0
G^2	892.957	922.212	929.808	959.190
AIC	1002.957	992.212	993.808	1007.190
corr(B1:ETA'U,Bi:ETA'U)	1.0	0.958	0.947	0.893
p-value for change in deviance test	1.0	0.083	0.034	0.0002

97d20

e 2) The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. The response was binary (700 ones and 300 zeroes) and logistic regression was used. The response plot for the full model B1 was good. The minimum AIC model had AIC = 990.134. Many of the predictors were factors, and the factor was counted as a predictor with a bad pvalue if all of the predictors in the factor had bad pvalues.

Which two models are the best candidates for the final submodel? Explain briefly why each of the other 2 submodels should not be used.

$\frac{300}{10} = 30$

B2 and B3

B1 has too many predictors

B4 has AIC > AIC(min) + 7 = 997.134

and pval is too low

14

e 3) Suppose that there are 14 subjects for a clinical trial. Five will get treatment A and nine will get treatment B. The names of the 14 people are listed below.
1 Reuben, 2 Mohammad, 3 Charles, 4 Mi-Ran, 5 Erin, 6 Chad, 7 Stephanie, 8 Yu, 9 Sankalpanand, 10 Buddika, 11 Joshua, 12 Cemile, 13 Kursad, and 14 Xiuquan.

Using the following computer output (randomization), write down the names of the subjects who receive treatment A.

```
> sample(1:14,5)
[1]  9  1 13 12  3
```

Sankalpanand, Reuben, Kursad, Cemile, Charles

14

or Reuben Charles Sankalpanand Cemile and Kursad

Math 282, Spring 2002 Final

		No Shot	One Shot	Two Shots	total
Flu	observed	24	9	13	46
	expected	(14.40)	(5.01)	(26.59)	
	cell chisq	[6.404]	[3.169]	[6.944]	
No Flu	observed	289	100	565	954
	expected	( )	(103.99)	(551.41)	
	cell chisq	[ ]	[0.153]	[ ]	
total		313	109	578	1000

4) The table above is a SRS of 1000 people from a community hospital. A new flu vaccine was provided free of charge in a two shot sequence over a period of two weeks. Some people received the two shot sequence, some appeared for one shot, and others received neither. It is desired to test whether there was any relationship between the number of flu shots and whether a person gets the flu.

a) Find the value of the expected count that is not given in the table. Find the 2 cell chi square contributions that need to be computed. Show work.

$\text{exp} = \frac{(\text{row tot})(\text{col tot})}{\text{tot}} = \frac{954\,(313)}{1000} = 298.602$

$\text{cell}\,\chi^2 = \frac{O-E}{E} = \frac{(289-298.602)^2}{298.602} = 0.3088$

$\frac{(565-551.41)^2}{551.41} = 0.3349$

b) Do a 4 step of hypotheses. Show how the appropriate table is used.

i) H_0 there is no relationship between number of shots and flu
H_A there is a relationship

ii) $\chi^2 = 6.404 + \cdots + .3349 = 17.314$

iii) $df = (I-1)(J-1) = (2-1)(3-1) = 2$ | $.001$ | 13.82

so $p\text{-val} < .001$

iv) reject H_0 there is a relationship between number of shots and flu

30

Current terms: (Acacia Bark Habitat Shrubs Stags Stumps)

	df	Deviance	Pearson X2	k	AIC
a) Delete: Shrubs	145	127.998	102.566	6	139.998

Current terms: (Acacia Bark Habitat Stags Stumps)

	df	Deviance	Pearson X2	k	AIC
b) Delete: Stumps	146	129.887	103.465	5	139.887

Current terms: (Acacia Bark Habitat Stags)

	df	Deviance	Pearson X2	k	AIC
c) Delete: Acacia	147	131.644	103.281	4	139.644

$\leftarrow J_{min} = J_I$

Current terms: (Bark Habitat Stags)

	df	Deviance	Pearson X2	k	AIC
d) Delete: Bark	148	138.685	110.187	3	144.685

Current terms: (Habitat Stags)

	df	Deviance	Pearson X2	k	AIC
e) Delete: Stags	149	149.861	123.14	2	153.861

5) a) Based on the output above, which model, a)–e), should be the first submodel to be examined. Explain briefly.

c) $J_I = J_{min}$

E3026

or -9

b) What are the predictors in this model?

Bark, Habitat, Stags

28

cholesterol level	CHD	no CHD
≥ 250	57	305
< 250	71	1098

6) The above tabled data is from a study on the association of the cholesterol level on the presence of coronary heart disease (CHD).

a) Find the odds ratio (for CHD versus no CHD).

$$\hat{\theta} = \frac{57\,(1098)}{71\,(305)} = 2.8901$$

b) What is the estimated relative risk (for CHD versus no CHD)?

$$\frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{57/(57+305)}{71/(71+1098)} = \frac{.15746}{.06074} = 2.5924$$

2.8901 − 8

c) Find the 95% CI for $\log(\theta)$.

$$\log\hat{\theta} \pm 1.96\,SE[\log\hat{\theta}] =$$

$$\log(2.8901) \pm 1.96\sqrt{\frac{1}{57}+\frac{1}{71}+\frac{1}{305}+\frac{1}{1098}}$$

$$= 1.0613 \pm 1.96\,(0.1893) = 1.0613 \pm 0.3709$$

$$= [(0.6904\ ,\ 1.4322)] = (L, U)$$

d) Exponentiate the endpoints of the interval in c) to give a 95% CI for θ.

$$(e^L, e^U) = [(1.9945\ ,\ 4.1879)]$$

56

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.901459	0.186877	10.175	0.000
t	0.556003	0.045780	12.145	0.000
t^2	-0.021346	0.002659	-8.029	0.000

7) Use the above output is for the $Y = number$ of new AIDS cases in Belgium from 1981 to 1993. The variable $x_1 = t = time$ was coded as 1 for 1981, 2 for 1982, ..., 13 for 1993. The variable $x_2 = t^2$. The output is from a Poisson regression. There were $n = 13$ cases.

a) Predict $\hat{\mu}(\boldsymbol{x})$ if $t = x_1 = 7.0$ and $x_2 = t^2 = 49.0$.

$ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 1.901459 + .5560030(7) - .021346(49)$

$= 4.7475$ ← just ESP = 9

$\hat{\mu} = e^{ESP} = e^{4.7475} = 115.2957$

or $e^{4.747526} = 115.2987$

b) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$

$Z_{01} = 12.145$

$p\text{-val} = 0$

reject Ho time is needed in the RR model for number given $(t.me)^2$ is in the model.

-2LR

c) Find a 95% confidence interval for β_2.

$\hat{\beta}_2 \pm 1.96\, SE(\hat{\beta}_2) = -.021346 \pm 1.96(.002659)$

$= -.021346 \pm .00521$

$= (-0.02656, -0.01614)$

45

Response = y = low, Sequential Analysis of Deviance
All fits include an intercept.

Predictor	df	Total Deviance	df	Change Deviance
Ones	188	234.672		
ht	187	230.650	1	4.02213
lwt	186	221.142	1	9.50777
ptl	185	215.964	1	5.17829
{F}race	183	210.850	2	5.11341
smoke	182	204.898	1	5.95270

8) Consider a study on whether the birthweight of a newborn baby is low or normal. Suppose that the response variable Y = low (0 if normal > 2500g, 1 if low birth weight < 2500 g). Predictors are ht = history of hypertension (0 no, 1 yes), lwt = weight of mother at last menstual period, ptl = history of premature labor (0 none, 1 one, 2 two, etc), factor race (1 white, 2 black, 3 other), and smoke during pregnancy (0 no, 1 yes). Use the above output to perform a 4 step deviance test.

i) H_0 $\beta = 0$ H_A $\beta \neq 0$

ii) $G^2(0|F) = 234.672 - 204.898 = 29.774$

iii) $df = 188 - 182 = 6$ | .001 | 22.46

pval $= 0 < .001$

iv) reject Ho there is a LR relationship between low and the predictors ht, ..., smoke.

9) Consider loglinear models in X, Y and Z. Then the saturated model (XYZ) is

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

Write model (XY, YZ) in terms of μ and λ's.

$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$

or-4

```
Output for Full Model, Response = nodal involvement,
Terms = (acid age grade size xray)
Number of cases: 53, Degrees of freedom: 47, Deviance: 48.126

Logistic Regression Output for Reduced Model,
Response = nodal involvement, Terms = (acid size xray)
Label       Estimate        Std. Error     Est/SE    p-value
Constant  -3.57564          1.18002        -3.030     0.0024
acid       2.06294          1.26441         1.632     0.1028
size       1.75556          0.738348        2.378     0.0174
xray       2.06178          0.777103        2.653     0.0080

Number of cases: 53, Degrees of freedom: 49, Deviance: 50.660
```

10) Treatment for prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. Let the response variable $y = nodal\ involvement$ (0 for absence, 1 for presence). Let $x_1 = acid$ (serum acid phosphatase level), $x_2 = size$ (= tumor size: 0 for small, 1 for large) and $x_3 = xray$ (xray result: 0 for negative, 1 for positive).

a) Predict $\hat{\pi}(\boldsymbol{x})$ if $x_1 = acid = 0.65$, $x_2 = 0$ and $x3 = 0$.

$ESP = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = -3.57564 + 2.06294(0.65) + 0 + 0$

$= -2.2347$

$e^{ESP} = e^{-2.2347} = 0.1070$

$\hat{\pi} = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{0.1070}{1.1070} = \boxed{0.09668}$

b) The full model used the 5 predictors listed above to the right of Terms. Perform a 4 step change in deviance test to see if the reduced model can be used.

i) H_0 the reduced model is good H_A use the full model

ii) $G^2(R|F) = 50.660 - 48.126 = 2.534$

iii) $df = 49 - 47 = 5 - 3 = 2$ | $\frac{0.25}{2.77}$

$pval > 0.25$

iv) fail to reject H_0 the reduced model is good.

28

e 11) Consider a study where the response variable is the total number of species recorded on each of 29 islands in the Galápagos Archipelago. Predictors are *log(areanear)* = log(area of the closest island) and *log(endem)* = log(the number of endemic species). (An endemic species is one that was not introduced from elsewhere). Output is for negative binomial regression.

```
outf <- glm.nb(Y~log(endem) + log(areanear))
outn <- glm.nb(Y~1);  anova(outf,outn)
   2 x log-lik.   Test    df  LR stat. Pr(Chi)
1   -313.5273
2   -204.5694    1 vs 2   2   108.9579   0
```

a) Use the above output to perform the likelihood ratio test.

$H_0: B = 0$ $H_A: B \neq 0$

$X^2(OIF) = 108.9579$

pval = 0

reject Ho there is an NBR relationship between number of species and the predictors log(areanear) and log(endem)

```
outr <- glm.nb(Y~log(endem));  anova(outf,outr)
2 x log-lik.   Test    df  LR stat.    Pr(Chi)
-208.1833
-204.5694     1 vs 2    1   3.613852   0.05730025
```

b) Use the above output to perform the change in LR test.

Ho the reduced model is good Ha use the full model

$X^2(RIF) = 3.6139$

pval = 0.0573

fail to reject Ho, the reduced model is good

28