STAT 504 Take-home Final Exam Spring 2016

Instructions:

- The take-home Final Exam has 5 problems. It is worth 100 pts. Your work has to be your own – **no group work**. You may consult any textbook and all the class notes. Please *do not* consult with anyone but Professor Morton.
- Please plan wisely — *no extensions will be given*. The exam is available April 30 and due by May 6 at 11:59 pm. You have 48 hours from the time you access the exam to complete it. Please submit on the ANGEL drop box. You can make the total of 3 submissions in case you want to update a previously submitted file.
- Do not include SAS or R (or any other computer program) output. Rather, be selective and edit and summarize the information so that it can be read and digested by others.
- Attach your code at the end.

1. (20 pts) For the following statements, answer true (T) or false (F).
   (a) _____ Overdispersion and latent variables can be used to deal with unobserved but predictive variables in a model.
   (b) _____ The quasi-symmetry model requires that marginal distributions be equal.
   (c) _____ The adjacent-category logit model and proportional-odds cumulative-logit model are reparameterizations of each other and give equivalent conclusions.
   (d) _____ A loglinear model with a structural zero fit by adding an indicator function, and the same model with the same indicator function but a 7 in place of the structural zero, will give the same estimate for the rest of the parameters (those not corresponding to the indicator function).
   (e) _____ In a proportional-odds cumulative-logit model for a 2×5 table (where the second variable is ordinal), the odds ratio comparing levels 1 and 2 of the second variable is the same as that comparing levels 2 and 4.
   (f) _____ A negative binomial GLM is often used as an alternative to Poisson regression with an overdispersion parameter.
2. (20 pts) The following table depicts 735 equal-sized plots of land in the Pennsylvania forest. The plots were measured at two times, year 1 and year 10, to determine whether Birch, Oak, or Other was the dominant tree species in the plot.

Y		Year 10		
e		birch	oak	other
a	birch	201	122	32
r	oak	110	175	17
1	other	4	24	50

(a) Collapse the Oak and Other category into non-birch, apply Mc Nemar's test, and report your conclusion.

(b) Test for quasi-independence on the full (3 by 3) table, reporting parameter values and significance as well as goodness of fit.

(c) Test for quasi-symmetry on the full (3 by 3) table, reporting parameter values and significance as well as goodness of fit.

(d) Test (goodness-of-fit only) for marginal homogeneity on the full (3 by 3) table.

(e) What are your conclusions, and which model fits best?

3. (20 pts) A securities exchange is trying to understand its order flow during periods between trades (transactions). Below is a table that records the number of orders received during a sample of such periods. Milliseconds is the length of time between transactions; orders is the count of messages received during this period; Price Movement is whether the price moved up or down in the transaction preceeding the no-trade period, and Security is which of three securities is being considered.

Milliseconds	Orders	Price movement	Security
673	68	Up	A
539	15	Up	B
843	107	Up	C
974	8	Down	A
11	22	Down	B
350	41	Down	C

(a) Fit a Possion regression model with main effects only and report the fit model. Interpret the parameters.

(b) Report the goodness of fit.

(c) Treat the $(Down, B)$ cell as anomalous and re-run the analysis. Do your conclusions change?

(d) What are your conclusions?

4. (20 pts) Consider the $2 \times 3 \times 2$ table of counts

		X			
		X=1		X=2	
		Z=1	Z=2	Z=1	Z=2
Y	Y=1	28	84	7	77
	Y=2	67	20	0	9
	Y=3	93	74	46	23

(a) Perform a test for 3-way independence (complete independence) and report your conclusion (treat the zero as a sampling zero).

(b) Does the MLE for the independence model exist? Why or why not?

(c) Treating the zero as the sampling zero, find "the best" model that fits these data. Explain your choice of the model.

(d) Now treat the zero as a structural zero and write the equation explaining counts with the loglinear model (YZ, ZX). Your answer should be the abstract equation, i.e. have lambdas rather than fit numbers. Although, I will accept the estimated equations too for partial credit.

(e) Treating the zero as a structural zero, write the equation for the logistic regression with Z as the response and X and Y as predictors. Your answer should be the abstract equation, i.e. have lambdas rather than fit numbers. Although, I will accept the estimated equations too for partial credit.

5. (20 pts) Suppose you have four random variables X, Y, Z, W with $4, 2, 2, 3$ levels respsectively.

(a) When performing a likelihood ratio test comparing the loglinear models XY, YZ, ZW and XY, YZW, how many degrees of freedom should you use for ΔG^2?

(b) Write the odds ratio for the 2×2 conditonal table where $X = 1, 2$ and $Y = 1, 2$, given $Z = 2$, and $W = 1$. Interpret this odds ratio.

(c) Suppose you fit the loglinear model XY, XZ, YZ, W with dummy coding where the last-category-baseline convention is used to obtain parameters such as $\lambda, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_\ell^W, \lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}$. Write the equation for the log of the expected counts, that is $log(\mu_{ijkl})$ where $i = 3, j = 2, k = 2, l = 1$ in terms of the λ parameters.