STAT 659 — Final Exam **Name:** ______________________________

Summer 2015

INSTRUCTIONS FOR THE STUDENT:

1. You have exactly 2 hours to complete the exam.
2. There are 13 pages including this cover sheet and 5 pages of SAS output.
3. Each lettered part of a question is worth 7 points unless otherwise marked.
4. Please answer all questions.
5. Show all your work on the test booklet.
6. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions.
7. You may use a calculator that does not have the capability of phoning, texting, or accessing the internet and three $8\frac{1}{2} \times 11$ formula sheets (you may use both sides). Do not use the textbook or class notes.
8. Carry out tests at level 0.05 unless otherwise stated.
9. Be sure to clearly state the hypotheses, the test statistic and its value, and conclusion for all tests.

I attest that I spent no more than 2 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature______________________________

INSTRUCTIONS FOR PROCTOR:

(1) Record the time at which the student starts the exam: ____________________

(2) Record the time at which the student ends the exam: ____________________

(3) Immediately after the student completes the exam, please scan the exam to a .pdf file and have the student upload it to webassign.

(4) Collect all portions of this exam at its conclusion. Do not allow them to take any portion with them.

(5) Please keep these materials until August 15, at which time you may either dispose of them or return them to the student.

I attest that the student has followed all the INSTRUCTIONS FOR THE STUDENT listed above and that the exam was scanned into a pdf and uploaded to webassign in my presence:

Proctor's Signature______________________________

Some Chi-Squared Percentiles

	Right-Tail Probability			
df	0.100	0.050	0.025	0.010
1	2.71	3.84	5.02	6.63
2	4.61	5.99	7.38	9.21
3	6.25	7.81	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
6	10.64	12.59	14.45	16.81
7	12.02	14.07	16.01	18.48
8	13.36	15.51	17.53	20.09
9	14.68	16.92	19.02	21.67
10	15.99	18.31	20.48	23.21

Some Normal Percentiles

Right-Tail Probability			
0.100	0.050	0.025	0.010
1.282	1.645	1.960	2.326

1. We return to the study by Long (1990), who investigated models to relate the number of publications (`art`) produced by Ph. D. biochemists as a function of gender (`fem` = 1 if female), marital status (`mar` = 1 if married), number of children (`kid5` = $0, 1, 2,$ or 3), a numerical rating of the prestige of the institution where the biochemist obtained the Ph. D. (`phd`) ranging from 0.75 to 4.62, and the number of articles written by the biochemist's mentor for the Ph. D. in the last 3 years (`ment`) ranging from 0 to 77. Here we will use the response variable `pubs` which equals 0 when `art` = 0, equals 1 when `art` = 1, and equals 2 when `art` $\geq$ 2. A proportional odds cumulative logit model and a baseline-category logit model with `pubs`= 0 being the baseline were fit to the data. Use the accompanying SAS output to help you answer this question.

   (a) Using the baseline logit model, estimate $P($`pubs` $= 0)$ when `phd` = 1, `ment` = 10, for an unmarried male with no children.

   (b) Using the proportional odds model, estimate $P($`pubs` $= 2)$ when `phd` = 1, `ment` = 10, for an unmarried male with no children.

2. A study was carried out of British occupational mobility where the occupational status of fathers and sons were placed in a two-way table where the status ranges from 1 (lowest) to 4 (highest).

	Son's status			
Father's status	1	2	3	4
1	174	84	154	55
2	78	110	223	96
3	150	185	714	447
4	42	72	320	411

(a) We examine the two-way table that dichotomizes the row and column categories to lower status and higher status. Carry out a test of whether the proportions of fathers in the two categories were the same as the proportions of sons in the two categories.

	Son's status		
Father's status	1 or 2 (lower)	3 or 4 (higher)	Total
1 or 2 (lower)	446	528	974
3 or 4 (higher)	449	1892	2341
Total	895	2420	3315

(b) Using the data in a), estimate the marginal odds ratio for the son's being in the **higher status category** relative to the father's being in the **higher status category**. Then obtain the conditional maximum likelihood estimate for the corresponding conditional odds ratio for the subject specific model.

- Marginal odds ratio

- Conditional odds ratio

(c) Several models were fit to the data in the four-by-four table resulting in the deviances and degrees of freedom in the following table:

Model	Deviance	Degrees of Freedom
Independence	452.4	9
Quasi-independence	79.4	5
Symmetry	30.1	6
Ordinal quasi-symmetry	12.8	5
Marginal homogeneity	27.8	3
Quasi-symmetry	2.3	3

Discuss the fit of these models and then chose the most appropriate model.

(d) Using the information in the table in part (c), carry out tests for marginal homogeneity in two ways:

- Test the fit of the model labeled "Marginal homogeneity".

- Assuming the quasi-symmetry model holds, carry out a test for marginal homogeneity of the data.

3. A study on the relationship between the number of previous pregnancies (`prevpreg` $= 0, 1, 2, 3, 4$) and the quality of prenatal care (`precare` $= 1$, if inadequate; $= 2$, if intermediate; $= 3$, if adequate) was based on records from a set of 3729 live births in Washington, D.C., from 1980 to 1985. Various loglinear models for two-way tables with ordered row and column categories were fit to the data resulting in the following table:

Model	G^2	*df*
Independence	11.03	8
Linear by Linear	6.09	7
Row effects	2.43	4
Column effects	4.88	6
Row and column effects	1.38	3
Saturated	0.00	0

(a) Use the information in the table above to test for independence versus (i) a general alternative and (ii) an alternative that takes into account the ordinality of rows and columns.

i. General alternative

ii. Ordered alternative

(b) Use the information in the table above to determine the most appropriate model.

4. A random sample of 3688 applicants for vocational education programs in major northeastern school districts was obtained. Each student was classified according to program applied for (`program`, 3 programs), gender (`gender`), and whether or not the student was accepted (`accept`). Various loglinear models identified by the first letter of the explanatory variable were fit to the data. Use the following table to help you answer the first three parts of this problem.

Model	G^2	*df*
(A,G,P)	1990.9	7
(AG,P)	1861.0	6
(A,GP)	1269.2	5
(AP,G)	755.7	5
(AG,AP)	625.8	4
(AG,GP)	1139.3	4
(AP,GP)	34.0	3
(AG,AP,GP)	9.1	2
(AGP)	0.0	0

(a) Carry out a test of equal odds ratios between `accept` and `gender` for the three programs.

(b) Assuming that the homogeneous association model holds, carry out a test of partial association of `accept` and `gender`, controlling for `program`.

Model	G^2	*df*
(A,G,P)	1990.9	7
(AG,P)	1861.0	6
(A,GP)	1269.2	5
(AP,G)	755.7	5
(AG,AP)	625.8	4
(AG,GP)	1139.3	4
(AP,GP)	34.0	3
(AG,AP,GP)	9.1	2
(AGP)	0.0	0

(c) Based on the table of deviances, which model would you recommend? Be sure to justify your answer.

(d) Estimate the odds ratio between `accept` and `gender` for applicants to the plumbing program (`program=plumbing`) using: the homogeneous association model and the saturated model. SAS output for these models is provided.

- Homogeneous association model

- Saturated model

5. We return to the study of the labor force participation by women in the United States using data from 1976. The sample consisted of 753 white, married women between the ages of 30 and 60 years old. The response variable is `y` $= 1$ (`inLF`) if the wife was in the labor force and $= 0$ (`NOtinLF`) if the wife was not in the labor force. The predictors that were measured were `k5` = the number of children 5 years old or under, `k618` = the number of children from 6 to 18 years old, `age` of the wife in years, `wc` $= 1$ if the wife attended college and $= 0$ otherwise, `hc` $= 1$ if the husband attended college and $= 0$ otherwise, `lwg` = the logarithm of the wife's estimated wages during the previous year, and `inc` = the family's estimated income excluding the wife's income. A logistic regression model was fit to these data with `y` as the response and predictors `k5, k618, age, wc,hc, lwg, inc`.

   (a) In the second test, students noticed that the marginal model plot indicated that the linear term in `lwg` was not appropriate for the model. The variable `lwg` was centered at its mean using $\texttt{lwg1} = \texttt{lwg} - \overline{\texttt{lwg}}$. The quadratic term $\texttt{lwg2} = (\texttt{lwg} - \overline{\texttt{lwg}})^2$ was added to the model. The resulting fitted model showed no signs of model inadequacy. To find a simpler model, a backward elimination process was carried out, removing the least significant predictor at each step and stopping when all the remaining predictors had $p-$values < 0.0001. Use the following table to choose the most appropriate model. Explain your reasoning.

Model	Predictors	Deviance	*df*	AIC
1	`k5, k618, age, wc, hc, lwg1, lwg2,inc`	751.8	744	769.8
2	`k5, k618, age, hc, lwg1, lwg2, inc`	751.8	745	767.8
3	`k5, age, hc, lwg1, lwg2, inc`	752.9	746	766.9
4	`k5, age, lwg1, lwg2, inc`	754.0	747	766.0
5	`k5, age, lwg1, lwg2`	770.2	748	780.2

   (b) Consider Model 5 with the attached SAS output to answer this part of the problem. Estimate the probability of being in the labor force for a woman with `k5` $= 1$, `age` $= 30$,and $\texttt{lwg} = \overline{\texttt{lwg}}$.