

UNIVERSITY OF SWAZILAND

SUPPLEMENTARY EXAMINATION PAPER 2010

**TITLE OF PAPER : TOPICS IN STATISTICS
(CATEGORICAL DATA ANALYSIS AND
GENERALIZED LINEAR MODELS)**

COURSE CODE : ST 405

TIME ALLOWED : THREE (3) HOURS

REQUIREMENTS : CALCULATOR AND STATISTICAL TABLES

INSTRUCTIONS : ANSWER ANY FIVE QUESTIONS

Question 1

- a) Prove that for a 2×2 contingency table, the chi-square test for independence is given by

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

(12 Marks)

- b) Suppose that independent random variables Y_1, Y_2, \dots, Y_n follow binomial distributions, that is

$$Y_i \sim B(m_i, \pi_i), i = 1, 2, \dots, n, \quad \text{where } P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i-y_i}.$$

Show that the binomial distribution is a member of the exponential family and that natural canonical link function for this distribution in the context of generalised linear models is given by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

(8 Marks)

Question 2

- a) A sports equipment company has commissioned an advertising agency to develop an advertising campaign for one of its new products. They can choose between two particular television commercials, A and B . To aid them in their decision, an experiment is performed in which 200 volunteers are randomly assigned to view one of the two commercials, 100 being assigned to each. After seeing the commercial, each volunteer is asked to state whether they would consider buying the product, with the following results.

		Commercial	
		A	B
Purchase product	No	70	80
	Yes	30	20

Apply a chi-squared test to these data and comment on your results. What recommendations, if any, would you make to the sports manufacturer concerning the choice of commercial for the proposed advertising campaign?

(15 Marks)

- b) In an industry the maintenance department wants to investigate the number of repairs to be made to 4 different makes of compressors, which are used in the 3 areas (North, Centre and South). The data on compressor failures are as follows:

Compressor	North	Centre	South
1	17	17	12
2	11	9	13
3	11	8	19
4	14	7	28

They engage a consultant to analyse the data and make a recommendation.

What would be your recommendation to the maintenance department?

(20 Marks)

Question 3

An art gallery is due to celebrate its 50th anniversary in 2005. As part of its celebrations, it wishes to commission a new sculpture to be displayed in the gallery. To find a suitable sculpture, it decided to run a competition in which it invited local artists to submit designs. A panel of experts selected a short-list of three designs for the gallery to choose from. To assist in the final decision, the gallery conducted a survey in which a random sample of local adults were sent copies of the three designs and asked to indicate their preference. The replies received from male and female adults are given in the following table.

		Preferred design		
		A	B	C
Males		129	24	47
Females		126	44	55

Carry out a suitable analysis to test whether or not the preference of design is the same for males and females.

(20 Marks)

Question 4

Given the following table with a set of models with values of G^2 (Likelihood Ratio Criterion) and p -value which relate to;

- Defendant's race D : W (White), B (Black) = Z variable
- Victim's race V : W (White), B (Black) = Y variable
- Death Penalty P : Y (yes), N (No) = X variable

Model	G^2	p -value
(D,V,P)	137.9	0.001
(DV,P)	8.1	0.04
(VP,DV)	1.9	0.39
(DP,VP,DV)	0.70	0.40
(DVP)	0	---

- Select two models that provide the best fit for the data and state the reasons for your choice. (2 Marks)
- From the two models select the best model and derive its log-linear model and give reasons for selecting it. (6 Marks)

- c) If $n_{111}=19$, $n_{112}=0$, $n_{121}=11$, $n_{122}=6$, $n_{211}=132$, $n_{212}=9$, $n_{221}=52$, and $n_{222}=97$, for the best model chosen in b) calculate \hat{U} , $\hat{U}_{D(1)}$ and $\hat{U}_{DV(11)}$.
(12 Marks)

Question 5

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases.

A research team from the U.S. measured all wells in an area of Araizahar upazila and labelled them with their arsenic level as well as a characterization as “safe” or “unsafe”, depending on whether the arsenic level was above or below the national standard of 0.5 in units of hundreds of micrograms per litre.

People with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction. The amount of water needed for drinking is low enough that adding users to a well would not exhaust its capacity. The surface water in this area is contaminated, hence the desire to use deep wells.

A few years later the researchers returned to see who had switched wells and found that 57.5% of the 3020 households with unsafe wells had switched. The team performed a series of analyses to understand the factors predictive of well switching among users of unsafe wells.

Variables:

distnear = the distance to the nearest safe well
ed = years of education
as = arsenic levels (ug/L)
logas = log-arsenic
edcXdistnc = (ed-med)*(distnear-mdistn)
med = mean of ed
mdist = mean of distnear

Model 1

Number of obs	=	3020				
LR chi2(3)	=	239.95				
Prob > chi2	=	0.0000				
Log Likelihood = -1939.077						
<hr/>						
switch	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
distnearest	-.0097893	.0010616	-9.22	0.000	-.0118699	-.0077087
logas	.888925	.068873	12.91	0.000	.7539365	1.023913
ed	.0431016	.0096435	4.47	0.000	.0242007	.0620024
_cons	-3.776544	.3315441	-11.39	0.000	-4.426358	-3.126729

Model 2

Number of obs	=	3020
LR chi2(4)	=	253.48

Prob > chi2 = 0.0000
 Log likelihood = -1932.3102

switch	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
distnearest	-.0100785	.0010827	-9.31	0.000	-.0122005 -.0079565
logas	.9046483	.069253	13.06	0.000	.7689149 1.040382
ed	.0441139	.0096605	4.57	0.000	.0251797 .063048
edcxdistnc	.0009318	.0002568	3.63	0.000	.0004284 .0014351
_cons	-3.843047	.333031	-11.54	0.000	-4.495775 -3.190318

- a) Define the two models and their hypotheses. Also, justify the appropriateness of the modelling procedure used. (6 Marks)
- b) How does model 1 compare with the models 2 in terms of parsimony and goodness of fit? (6 Marks)
- c) Explain the effect of education on well-switching, including how it varies for different types of respondents. (8 Marks)

Question 6

In the data from the US 1996 General Social Survey, say they were primarily interested in volunteer, a variable representing the number of volunteer activities in the past year. Note that gender is a dummy for females best called **female** and race is a dummy for non-whites best called **nonwhites**. Two other predictors of interest are education and income. A GLM was fitted and results are:

Number of obs = 1944
 LR chi2(7) = 121.96
 Prob > chi2 = 0.0000
 Log likelihood = -1675.116

volteer	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
female	.2071766	.0843299	2.46	0.014	.0418931 .3724602
nonwhite	-.6738627	.2013554	-3.35	0.001	-1.068512 -.2792134
nonwfemale	.6123789	.239987	2.55	0.011	.1420129 1.082745
educate	.1250034	.0189202	6.61	0.000	.0879206 .1620862
educatecsq	-.0131087	.0048738	-2.69	0.007	-.0226612 -.0035562
income	.1054104	.0270514	3.90	0.000	.0523906 .1584302
incomecsq	.0112797	.0048598	2.32	0.020	.0017547 .0208048
_cons	-3.944508	.3626887	-10.88	0.000	-4.655364 -3.233651

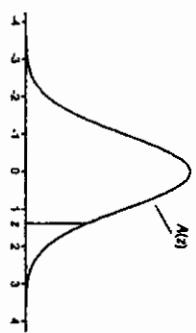
- a) Interpret the coefficients and comment briefly on their significance on the basis of the Wald test. (Note that we could do likelihood ratio tests but we'll stick to Wald tests for simplicity.) (5 Marks)
- b) Check if it was appropriate to treat education as a linear effect by introducing a quadratic term and testing its significance. (5 Marks)
- c) Verify that you also need to introduce a quadratic term for income. (Working with log-income doesn't help in this case.) (5 Marks)
- d) Test whether the female effect differs by ethnicity and interpret carefully your estimated coefficients.

STATISTICAL TABLES

Cumulative Standardized Normal Distribution

Table A.1

$\Phi(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of $\Phi(z)$ of particular importance:



z	$\Phi(z)$
-1.645	0.9500
0.960	0.9750
2.326	0.9900
2.376	0.9950
3.090	0.9990
3.291	0.9993

Lower limit of right 5% tail
Lower limit of right 2.5% tail
Lower limit of right 1% tail
Lower limit of right 0.5% tail
Lower limit of right 0.1% tail
Lower limit of right 0.05% tail

- Cumulative normal distribution**
- Critical values of the t distribution**
- Critical values of the F distribution**
- Critical values of the chi-squared distribution**

Table A.2
t Distribution: Critical Values of t

Degrees of freedom	Two-tailed test:	Significance level									
		10%	5%	2%	1%	0.5%	0.2%	0.1%	0.05%	0.02%	0.01%
1	6.316	12.766	31.821	63.657	318.309	636.619	193.35	19.37	18.38	19.40	19.41
2	2.920	4.303	6.863	9.925	22.327	31.599	12.524	4.297	4.541	4.609	4.641
3	2.353	3.182	4.541	5.841	10.215	12.524	5.717	6.10	6.40	6.60	6.74
4	2.132	2.776	3.747	4.604	8.610	8.629	4.031	4.893	4.95	4.98	4.99
5	2.015	2.571	3.385	4.031	7.076	7.173	4.031	4.893	4.95	4.98	4.99
6	1.943	2.447	3.143	3.707	5.208	5.959	4.118	4.874	4.93	4.96	4.97
7	1.894	2.365	2.998	3.499	4.783	5.408	3.844	4.594	4.85	4.88	4.91
8	1.860	2.306	2.896	3.353	4.501	5.041	3.609	4.359	4.61	4.64	4.67
9	1.833	2.263	2.821	3.250	4.297	4.781	3.577	4.327	4.58	4.61	4.64
10	1.812	2.228	2.764	3.169	4.144	4.597	3.459	4.214	4.47	4.50	4.53
11	1.796	2.190	2.718	3.053	4.023	4.437	3.306	4.053	4.31	4.34	4.37
12	1.782	2.179	2.691	3.023	3.930	4.318	3.274	3.923	4.18	4.21	4.24
13	1.771	2.160	2.650	3.012	3.893	4.221	3.220	3.869	4.14	4.17	4.20
14	1.761	2.145	2.624	2.977	3.787	4.140	3.166	3.823	4.11	4.14	4.17
15	1.753	2.131	2.602	2.947	3.783	4.073	3.135	3.783	4.07	4.10	4.13
16	1.746	2.120	2.583	2.921	3.696	4.019	3.106	3.755	4.06	4.09	4.12
17	1.740	2.110	2.567	2.898	3.666	3.983	3.075	3.725	4.02	4.05	4.08
18	1.734	2.101	2.552	2.876	3.610	3.922	3.055	3.705	4.01	4.04	4.07
19	1.729	2.093	2.539	2.861	3.583	3.883	3.035	3.694	3.91	3.94	3.97
20	1.725	2.086	2.538	2.845	3.551	3.850	3.015	3.674	3.89	3.92	3.95
21	1.721	2.080	2.518	2.821	3.527	3.819	2.994	3.654	3.81	3.84	3.87
22	1.717	2.074	2.509	2.819	3.505	3.792	2.974	3.634	3.78	3.81	3.84
23	1.714	2.069	2.500	2.807	3.485	3.768	2.954	3.614	3.73	3.76	3.79
24	1.711	2.064	2.492	2.797	3.467	3.745	2.934	3.594	3.69	3.72	3.75
25	1.708	2.060	2.485	2.787	3.450	3.723	2.914	3.574	3.67	3.70	3.73
26	1.706	2.056	2.479	2.779	3.433	3.707	2.894	3.554	3.66	3.69	3.72
27	1.703	2.052	2.473	2.771	3.421	3.690	2.874	3.534	3.64	3.67	3.70
28	1.701	2.048	2.467	2.763	3.408	3.674	2.854	3.514	3.62	3.65	3.68
29	1.699	2.045	2.462	2.756	3.396	3.659	2.834	3.494	3.61	3.64	3.67
30	1.697	2.042	2.457	2.750	3.383	3.646	2.814	3.474	3.59	3.62	3.65
32	1.694	2.037	2.449	2.738	3.365	3.622	2.794	3.454	3.57	3.60	3.63
34	1.691	2.031	2.441	2.728	3.348	3.601	2.765	3.434	3.55	3.58	3.61
36	1.688	2.028	2.434	2.719	3.333	3.582	2.736	3.414	3.53	3.56	3.59
38	1.686	2.024	2.429	2.712	3.319	3.566	2.707	3.395	3.51	3.54	3.57
40	1.684	2.021	2.423	2.704	3.307	3.551	2.678	3.376	3.49	3.52	3.55
42	1.682	2.018	2.418	2.698	3.296	3.538	2.651	3.357	3.47	3.50	3.53
44	1.680	2.014	2.414	2.693	3.286	3.526	2.624	3.343	3.45	3.48	3.51
46	1.679	2.011	2.411	2.687	3.277	3.515	2.596	3.324	3.43	3.46	3.49
48	1.677	2.007	2.407	2.682	3.269	3.505	2.567	3.313	3.42	3.45	3.48
50	1.676	2.005	2.403	2.676	3.261	3.496	2.538	3.294	3.41	3.44	3.47
52	1.671	2.000	2.399	2.669	3.253	3.489	2.509	3.275	3.39	3.42	3.45
54	1.667	1.994	2.384	2.661	3.243	3.481	2.480	3.254	3.37	3.40	3.43
56	1.664	1.990	2.374	2.659	3.239	3.475	2.451	3.234	3.35	3.38	3.41
58	1.662	1.987	2.368	2.652	3.236	3.469	2.422	3.213	3.33	3.36	3.39
60	1.660	1.984	2.364	2.656	3.174	3.460	2.393	3.193	3.31	3.34	3.37
62	1.659	1.980	2.358	2.647	3.160	3.373	2.364	3.173	3.29	3.32	3.35
64	1.658	1.976	2.351	2.645	3.159	3.377	2.335	3.144	3.26	3.29	3.32
66	1.653	1.971	2.345	2.609	3.131	3.340	2.306	3.115	3.23	3.26	3.29
68	1.650	1.968	2.339	2.592	3.118	3.323	2.277	3.086	3.19	3.22	3.25
70	1.649	1.966	2.336	2.588	3.111	3.315	2.248	3.057	3.16	3.19	3.22
72	1.648	1.963	2.334	2.586	3.107	3.310	2.219	3.028	3.13	3.16	3.19
74	1.647	1.964	2.333	2.584	3.104	3.307	2.190	3.009	3.12	3.15	3.18
76	1.644	1.960	2.326	3.090	3.291	3.291	2.161	2.981	3.09	3.12	3.15

Table A.3
F Distribution: Critical Values of F (#% significance level)

n ₁	n ₂	Significance level															
		1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	
1	1	161.43	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.72	248.01	
2	3	18.51	19.00	19.16	19.23	19.30	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45		
3	4	10.13	9.53	9.28	9.12	9.01	8.94	8.89	8.83	8.81	8.79	8.71	8.69	8.67	8.66		
4	5	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.98	5.91	5.87	5.84	5.81	5.80		
5	6	6.65	5.79	5.41	5.19	5.03	4.93	4.85	4.77	4.69	4.62	4.57	4.52	4.48	4.44		
6	7	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.89		
7	8	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47		
8	9	5.12	4.26	3.86	3.53	3.37	3.29	3.21	3.18	3.14	3.07	3.03	2.99	2.95	2.94		
9	10	4.86	4.10	3.71	3.48	3.28	3.17	3.09	3.02	2.97	2.91	2.86	2.81	2.77	2.73		
10	11	4.58	3.81	3.49	3.28	3.08	2.97	2.89	2.82	2.75	2.69	2.64	2.59	2.54	2.50		
11	12	4.31	3.61	3.31	3.12	2.92	2.82	2.74	2.67	2.60	2.54	2.49	2.44	2.39	2.34		
12	13	4.07	3.37	3.17	2.98	2.78	2.68	2.60	2.52	2.45	2.39	2.34	2.29	2.24	2.19		
13	14	3.84	3.14	2.94	2.75	2.55	2.45	2.37	2.29	2.21	2.15	2.10	2.05	2.00	1.95		
14	15	3.64	2.94	2.74	2.54	2.34	2.24	2.16	2.08	2.00	1.95	1.90	1.85	1.80	1.75		
15	16	3.45	2.75	2.55	2.35	2.15	2.05	1.97	1.89	1.81	1.76	1.71	1.66	1.61	1.56		
16	17	3.27	2.57	2.37	2.17	1.97	1.87	1.79	1.71	1.63	1.58	1.53	1.48	1.4			

Table A.3 (continued)

TABLE A.3 (continued)

TABLE 3 (continued)

F Distribution: Critical Values of F (1% significance level)

Table A.3 (continued)

v_1	25	30	35	40	50	60	70	75	100	150	200
1	4.24003	4.24003	4.24003	4.24003	4.24003	4.24003	4.24003	4.24003	4.24003	4.24003	4.24003
2	3.99936	3.99947	3.99947	3.99947	3.99947	3.99947	3.99947	3.99947	3.99947	3.99947	3.99947
3	3.12544	3.12545	3.12545	3.12545	3.12545	3.12545	3.12545	3.12545	3.12545	3.12545	3.12545
4	4.45702	4.45743	4.45743	4.45743	4.45743	4.45743	4.45743	4.45743	4.45743	4.45743	4.45743
5	2.938	2.947	2.947	2.947	2.947	2.947	2.947	2.947	2.947	2.947	2.947
6	16.535	16.657	16.654	16.644	16.631	16.621	16.612	16.603	16.593	16.589	16.589
7	12.659	12.733	12.741	12.733	12.720	12.711	12.704	12.695	12.687	12.682	12.682
8	10.256	10.111	10.000	9.922	9.840	9.753	9.655	9.577	9.499	9.455	9.455
9	8.659	8.554	8.446	8.337	8.256	8.179	8.104	8.026	7.953	7.896	7.896
10	7.60	7.47	7.37	7.30	7.19	7.13	7.05	6.98	6.91	6.87	6.87
11	6.811	6.68	6.59	6.52	6.42	6.33	6.28	6.21	6.14	6.10	6.10
12	6.22	6.09	6.00	5.93	5.85	5.76	5.69	5.62	5.56	5.52	5.52
13	5.755	5.63	5.54	5.47	5.37	5.29	5.24	5.17	5.10	5.07	5.07
14	5.318	5.23	5.17	5.10	5.00	4.93	4.87	4.81	4.74	4.71	4.71
15	5.07	4.95	4.86	4.70	4.64	4.57	4.51	4.44	4.41	4.37	4.37
16	4.832	4.70	4.61	4.54	4.45	4.39	4.32	4.26	4.19	4.16	4.16
17	4.630	4.49	4.40	4.33	4.24	4.18	4.11	4.05	3.98	3.95	3.95
18	4.42	4.20	4.12	4.06	3.98	3.91	3.87	3.80	3.77	3.70	3.70
19	4.26	4.14	4.06	3.99	3.90	3.84	3.78	3.71	3.65	3.61	3.61
20	4.12	4.00	3.92	3.86	3.77	3.70	3.64	3.58	3.51	3.48	3.48
21	4.00	3.88	3.80	3.74	3.64	3.58	3.52	3.46	3.39	3.36	3.36
22	3.89	3.78	3.70	3.63	3.54	3.48	3.41	3.35	3.28	3.23	3.23
23	3.79	3.68	3.60	3.53	3.44	3.38	3.32	3.25	3.19	3.16	3.16
24	3.71	3.59	3.51	3.45	3.36	3.29	3.23	3.17	3.10	3.07	3.07
25	3.63	3.52	3.43	3.37	3.28	3.21	3.15	3.09	3.03	2.99	2.99
26	3.56	3.44	3.36	3.30	3.21	3.15	3.08	3.02	2.95	2.92	2.92
27	3.49	3.38	3.20	3.13	3.04	3.00	3.02	2.96	2.89	2.86	2.86
28	3.43	3.32	3.24	3.16	3.09	3.02	2.96	2.90	2.83	2.80	2.80
29	3.38	3.27	3.18	3.12	3.03	2.97	2.91	2.84	2.78	2.74	2.74
30	3.33	3.22	3.13	3.07	2.98	2.92	2.86	2.79	2.73	2.69	2.69
35	3.13	3.02	2.93	2.87	2.78	2.72	2.66	2.59	2.52	2.49	2.49
40	2.98	2.87	2.79	2.73	2.64	2.57	2.51	2.44	2.38	2.34	2.34
50	2.79	2.68	2.60	2.53	2.44	2.38	2.31	2.23	2.14	2.10	2.10
60	2.57	2.45	2.47	2.41	2.33	2.24	2.19	2.11	2.05	2.01	2.01
70	2.38	2.27	2.19	2.13	2.05	2.00	1.94	1.89	1.82	1.79	1.79
80	2.47	2.36	2.27	2.21	2.11	2.06	1.96	1.89	1.85	1.81	1.81
90	2.42	2.31	2.26	2.16	2.10	2.05	1.96	1.89	1.85	1.81	1.81
100	2.43	2.32	2.24	2.17	2.08	2.01	1.94	1.87	1.83	1.79	1.79
150	2.37	2.26	2.18	2.11	2.02	1.95	1.88	1.81	1.73	1.68	1.68
200	2.32	2.21	2.12	2.06	1.96	1.89	1.82	1.74	1.66	1.62	1.62
250	2.25	2.13	2.03	1.97	1.87	1.80	1.77	1.65	1.56	1.51	1.51
300	2.21	2.10	2.01	1.94	1.85	1.78	1.70	1.62	1.53	1.48	1.48
400	2.18	2.07	1.98	1.92	1.82	1.75	1.67	1.59	1.50	1.43	1.43
500	2.17	2.05	1.97	1.90	1.80	1.73	1.67	1.59	1.50	1.43	1.43
600	2.16	2.04	1.96	1.89	1.79	1.71	1.64	1.56	1.46	1.41	1.41
700	2.15	2.03	1.93	1.88	1.78	1.71	1.63	1.55	1.46	1.40	1.40
1000	2.14	2.02	1.94	1.87	1.77	1.69	1.62	1.53	1.44	1.38	1.38

Table A.4

Degrees of freedom	Significance level
1	3.841
2	5.991
3	7.815
4	9.488
5	11.070
6	12.592
7	14.067
8	15.507
9	16.919
10	21.666
15	26.124
20	29.877
25	30.513
30	30.507
35	30.209
40	29.588