UNIVERSITY OF SWAZILAND

SUPPLEMENTARY EXAMINATION PAPER 2016

TITLE OF PAPER	: TOPICS IN STATISTICS (STATISTICAL MODELLING)
COURSE CODE	: ST 405
TIME ALLOWED	: TWO (2) HOURS
REQUIREMENTS	: CALCULATOR AND STATISTICAL TABLES
INSTRUCTIONS	: ANSWER ANY FIVE QUESTIONS

Question 1

A marketing research firm was engaged by an automobile manufacturer to conduct a pilot study to examine the feasibility of using logistic regression for ascertaining the likelihood that a family will purchase a new car during the next year. A random sample of 33 suburban families was selected. Data on annual family income and the current age of the oldest family automobile were obtained. A follow-up interview conducted 12 months later was used to determine whether the family actually purchased a new car or did not purchase a new car. The model in Appendix I was fitted;

a) State the response function.

(3 Marks)

b) Using the output in Appendix I (coefficients) advise appropriately.

(10 Marks)

c) What is the estimated probability that a family with annual income of E50,000 and an oldest car of 3 years will purchase a new car next year?

(3 Marks)

d) Using the output Appendix II, state whether the two-factor interaction effect between annual family income and age of oldest automobile should be added to the regression model containing family income and age of oldest automobile as first-order terms; use $\alpha = 0.05$. What is the approximate p-value?

(4 Marks)

Question 2

In the 2002 Winter Olympic Games held at Salt Lake City there was concern that figure skating judges may have judged with bias for certain skaters according to geopolitical preferences. Consider Table B which presents the results of the 9 judges of the "long program" in women's figure skating. All judges rated these two skaters first or second; the first place rating, or preferred skater, is reported in the table. Each judge is placed into a region based upon her/his country of origin: EE refers to Belarus, Russia and Slovakia; WE refers to Denmark, Finland, Italy, and Germany; and NA refers to Canada and the United States.

TABLE B. Number of First place ratings

	Figure Skater		
Region	Slutskaya	Hughes	Total
EE	3	0	3
WE	1	3	4
NA	0	2	2
Total	4	5	9

Briefly, summarize your results for a report to the International Olympic Committee (IOC). What do you conclude about the charge of geopolitically based bias in figure skating judging? Explain any assumptions of the statistical methods that led you to this conclusion (the IOC will consult a statistician to independently review your findings.)

(20 Marks)

Question 3

The table below displays a 4-way cross-classification of data related to complaints of symptoms of a respiratory disease, byssinosis, which occurs among textile mill workers.

Table. Frequency table of Byssinosis Complaints

WORKPLACE CONDITIONS	YEARS EMPLOYMENT	SMOKING	COMPLAINTS yes	COMPLAINTS no
Dusty	<10	yes	30	203
Dusty	<10	no	7	119
Dusty	>=10	yes	57	161
Dusty	>=10	no	11	81
Not Dusty	<10	yes	14	1340
Not Dusty	<10	no	12	1004
Not Dusty	>=10	yes	24	1360
Not Dusty	>=10	no	10	986

Create 2×2 tables stratified by years of employment and test for independence between Workplace conditions and complaints of byssinosis by comparing the Pearson Chi square test and Mantel- Haenszel Chi-Square test.

(20 Marks)

Question 4

a) If we write the probability density function (p.d.f.) for the GLM in the form

$$f(y_i|\theta_i,\emptyset) = exp\left[\frac{y_i\theta_i - m(\theta_i)}{h(\emptyset)}\right] + n(y_i,\emptyset)$$

then prove that $E(Y_i) = m'(\theta_i)$ and $Var(Y_i) = m''(\theta_i)\, h(\emptyset)$.

(10 Marks)

b) If the random variable Y_i follows a Gamma distribution, with scale parameter θ and shape parameter $\emptyset$, then it has a p.d.f.

$$f(y_i|\theta_i,\emptyset) = \frac{y_i^{\emptyset-1}\theta_i^{\emptyset}\, e^{-y_i\theta_i}}{\Gamma(\emptyset)}$$

Show that the distribution is a member of the exponential family, and find $E(Y_i)$ and $Var(Y_i)$.

(10 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
```

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 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)

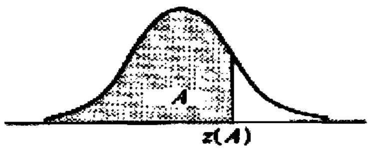
d) Test whether the female effect differs by ethnicity and interpret carefully your estimated coefficients.

(5 Marks)

Normal Distribution

Table C-1. Cumulative Probabilities of the Standard Normal Distribution.

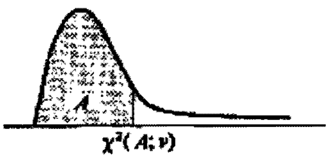
Entry is area A under the standard normal curve from $-\infty$ to $z(A)$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table C-2. Percentiles of the χ^2 Distribution

Entry is $\chi^2(A;\nu)$ where $P\{\chi^2(\nu) \le \chi^2(A;\nu)\} = A$

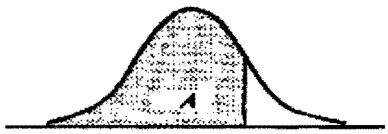


ν	A: .005	.010	.025	.050	.100	.900	.950	.975	.990	.995
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

Student's Distribution (t Distribution)

Table C-4 Percentiles of the t Distribution

Entry is $t(A;\ \nu)$ where $P\{t(\nu) \leq t(A;\ \nu)\} = A$



$t(A;\ \nu)$

ν	A .60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Table C-4 (Continued) Percentiles of the *t* Distribution

	A						
ν	.98	.985	.99	.9925	.995	.9975	.9995
1	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.849
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.291