**UNIVERSITY OF NAIROBI**

**UNIVERSITY EXAMINATION 2011/2012**

**2ND YEAR EXAMINATION FOR THE DEGREE OF MASTER OF SCIENCE IN SOCIAL STATISTICS**

**SCHOOL OF MATHEMATICS, CHIROMO CAMPUS**

**COURSE: STA 602: MODELLING AND ANALYSIS OF SOCIAL DATA**

**DATE: 4TH MARCH 2013: TIME: 4.00-6.00 PM**

**INSTRUCTIONS: ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS**

**QUESTION ONE (30 MARKS)**

a) A group research scientists conducted an experiment using mice with an aim of investigating whether or not exposure to cigarette smoke encourages the development of lung tumour. The binary observations are as recorded below. Study carefully and answer the subsequent questions

|  |  |  |
| --- | --- | --- |
| Group | Tumour present | Tumour absent |
| Treated | 21 | 2 |
| Control | 19 | 13 |

Compute

1. proportion of mice that develop a tumour in the treated group (1 mk)
2. the ratio of the odds of a tumour occurring in the targeted group relative to the control group. Interpret your results (3 mks)
3. the 95% confidence interval of the true log odds ratio (3 mks)
4. would you conclude that the odds of a tumour is greater amongst exposed group? Show your working (2 marks)

b) State the difference between the following data sources for social statistics

1. Surveys and census (3 mks)
2. Administrative records and civil registration system (3 mks)
3. State any three uses of social statistics (3 mks)

c) A data on housing price in central business district was obtained from Kisii municipality since 2000 to 2012. An investigator was interested to determine trend (T), cyclical component (C), seasonal component (E) and irregular component (A) based on the data. The computed variance and average for seasonal difference were 3,865.50 and 128.36 whereas that of seasonal quotient was 0.01 and 1.10, respectively. An investigator wanted to fit either the additive model of the form *Yt = Tt + Ct + Et + At* or multiplicative model of the form *Yt = Tt \* Ct \* Et \*At.*

1. which of these two models suited the data? show your workings (4 mks)
2. what is the difference between four parameters/components (T, C, E and A) in the model ? (4 mks)

d) Suppose X1, X2,…..XK is a vector of latent random variables representing the thematic content of a document in Gap model where each *xi* encodes the total length of passages about topic *i* in the document. Then *XI* is assumed to be an independent random variable with a gamma distribution of the form *p*(*x*) = {*x*(*a−*1)exp(*−x/b*)}/ *ba*Γ(*a*); where *ai* is the shape parameter and *bi* scale parameter. The corresponding probability density function of the random variable xi is

*p*(*xi*) = *ai*(*aixi/ci*)*ai−*1 exp(*−xiai/ci*)/*ci*Γ(*ai*) for *xi >* 0. Suppose *a1 = 0.5, a2=1, a3=2.5 a4 =3* and *a5 =4; b1=1, b2=2, b3=4, b4=5* and *b5=7* for corresponding random variable xi.

1. compute mean and standard deviation of the random variable x. comment on your answer (3 marks)
2. What kind of statistical distribution does the gamma distribution results when a=1. Illustrate the distribution (1 mark)

**QUESTION TWO (20 MARKS)**

a) A research scientist conducted a study on the effect of application of inorganic fertilizer (10mg/ha of nitrogen) on maize yield (t/ha). The data generated was subjected to generalized linear modeling procedures. The results are as shown below. Study them carefully to answer subsequent questions.

Regression analysis: Response variate: yld\_t\_ha; Fitted terms: Constant, inorg\_t\_ha

Source d.f. s.s. m.s. v.r. F pr.

Regression 1 22.46 22.4639 34.68 <.001

Residual 34 22.03 0.6478

Total 35 44.49 1.2711

**Parameter estimates**

Parameter estimate s.e. t(34) t pr.

Constant 1.228 0.366 3.36 0.002

inorg\_t\_ha 0.1419 0.0241 5.89 <.001

1. Compute total variability accounted by application of inorganic fertilizer (2 mks)
2. Formulate the model and interpret the parameters as appropriate (3 mks)
3. Compute the 95% confidence interval of the inorganic estimate (2 mks)
4. Interpret the implication of the 95% CI (2 mks)
5. Interpret the implication of model checking on the basis of the following graphics (4mks)



b) A baseline survey was taken in Coastal regions covering Kwale, Kilifi, Malindi, Bura, Lamu and Mariakani to assess the effect of seven technologies denoted as T1, T2, T3, T4, T5, T6 and T7 on maize yield in farmers’ field. The data collected was analyzed using using linear mixed model of the form .

The model output was as below. Study it carefully and answer the corresponding questions

\*\*\* Estimated Variance Components \*\*\*

Random term Component S.e.

SITE 0.2516 0.0992

SITE.FARM 0.3535 0.0616

\*\*\* Residual variance model \*\*\*

Term Factor Model(order) Parameter Estimate S.e.

Residual Identity Sigma2 0.475 0.0325

\*\*\* Wald tests for fixed effects \*\*\*

Fixed term Wald statistic d.f. Wald/d.f. Chi-sq prob

\* Sequentially adding terms to fixed model

Technologies 663.07 11 60.28 <0.001

2.455 Standard error: 0.1165

\*\*\* Table of predicted means for technologies \*\*\*

Technologies T1 T2 T3 T4 T5 T6 T 7

1.308 2.984 2.681 3.369 1.495 3.377 1.858

Standard error of differences:

Average 0.1510 Maximum 0.1585 Minimum 0.1457

Average variance of differences: 0.02281

1. Interpret the random variance components (2 mks)
2. The model output indicated that there were statistically significant differences among technologies promoted. Which of these technologies differed? Show your working (5 marks)

**QUESTION THREE (20 MARKS)**

a) Table below shows a subset of data on average prices per square metre of houses in Mombasa Countydistrict taken quarterly from 2005 to 2007. Study it carefully and answer subsequent questions.

|  |  |  |  |
| --- | --- | --- | --- |
| **Quarter/year** | **2005** | **2006** | **2007** |
| **Quarter 1(Q1)** | 857 | 994 | 1148 |
| **Q2** | 891 | 1030 | 1193 |
| **Q3** | 926 | 1065 | 1254 |
| **Q4** | 953 | 1096 | 1287 |

Calculate:

1. Price quotients (4 marks)
2. Seasonal differences (4 marks)
3. Comment on your results (2 marks)

b) Time series model (Autoregressive Integrated Moving average) of the following forms were fitted to the whole data set on average price per square metre. The output was summarized in the table below. Study it carefully and answer the following questions.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ARIMA model with | Parameters coefficient | s.e | Z | p-value | Log-likelihood ratio | Akaike criterion |
| AR only | 0.543 | 0.123 | 1.999 | 0.000 | -43.54 | 88.02 |
| AR and MA | 0.214 | 0.762 | 0.654 | 0.341 | -47.25 | 97.30 |
|  | 0.383 | 0.620 | 0.619 | 0.768 |  |  |
| AR, Difference and MA | 0.373 | 0.426 | 3.072 | 0.039 | 37.21 | 62.49 |
|  | -1.000 | 0.165 | -4.107 | <0.001 |  |  |

1. Which of the ARIMA models fitted the data better and why? (4 marks)
2. Which of the ARIMA model had poorly estimated coefficients and why? (3 marks)
3. What is the implication of such poor estimates? (3 marks)

**QUESTION FOUR (20 MARKS)**

1. Suppose that the education system is denoted by the integers 1, 2, 3……N, where N is the number of possible states of the education system and assume that the system is time homogenous so that there is a fixed probability *Pi*j, that a student in state *i* at time *(t-1)* will transfer to state *j* at time *t*. Assume further that the system has r absorbing and s non-absorbing states, such that r+s =N. This will then give rise to transition matrix of the canonical form of the Markovian model,

P = [I O]

[ R Q]

Define the following components of the transition matrix

1. I (1 mark)
2. O (1 mark)
3. R (2 marks)
4. Q (2 marks)
5. The model was fitted on the primary school system in Kenya for seven grades to obtain the school staying ratio, drop out, completion ratios, expected length of schooling and survival time. The extracted results are as presented in Table below. Use the output in each of the tables to answer questions that follow.

Table on fraction of pupils now in grade (standard) one who n years later will be in standard j, j=2. 3,….7; n= 1, 2,…8

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grade  Year | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | Boys (B) | 0.1346 | 0.6588 |  |  |  |  |  |
| Girls (G) | 0.1280 | 0.6594 |  |  |  |  |  |
| 2 | B | 0.0181 | 0.1784 | 0.5458 |  |  |  |  |
| G | 0.0164 | 0.1723 | 0.5507 |  |  |  |  |
| 3 | B | 0.0024 | 0.0362 | 0.2187 | 0.4616 |  |  |  |
| G | 0.0021 | 0.0338 | 0.2154 | 0.4715 |  |  |  |
| 4 | B | 0.0003 | 0.0065 | 0.0584 | 0.2448 | 0.3789 |  |  |
| G | 0.0003 | 0.0059 | 0.0562 | 0.2479 | 0.3891 |  |  |
| 5 | B |  | 0.0011 | 0.0130 | 0.0812 | 0.2490 | 0.3246 |  |
| G |  | 0.0010 | 0.0122 | 0.0815 | 0.2572 | 0.3226 |  |
| 6 | B |  | 0.0002 | 0.0026 | 0.0215 | 0.0982 | 0.2622 | 0.2538 |
| G |  | 0.0002 | 0.0024 | 0.0214 | 0.1020 | 0.2670 | 0.2196 |
| 7 | B |  |  | 0.0005 | 0.0050 | 0.0301 | 0.1236 | 0.2468 |
| G |  |  | 0.0004 | 0.0049 | 0.0315 | 0.1291 | 0.2147 |
| 8 | B |  |  | 0.0001 | 0.0011 | 0.0079 | 0.0444 | 0.1373 |
| G |  |  | 0.0001 | 0.0010 | 0.0083 | 0.0476 | 0.1200 |

1. Compute the school staying ratio of boys and girls from year one to year eight (4 marks).
2. Comment on your results of school staying ratio between boys and girls (3 marks)
3. What is the probability of the pupils now in standard one will be in standard:
4. four next four years ( 1 mark)
5. six in the next five years (1 mark)
6. seven the next eight years (1 mark)
7. provide your observation on the transitional probabilities for A, B and C ( 2 marks)
8. What is the significance of these results and model in educational planning? (2 marks)

**QUESTION FIVE (20 MARKS)**

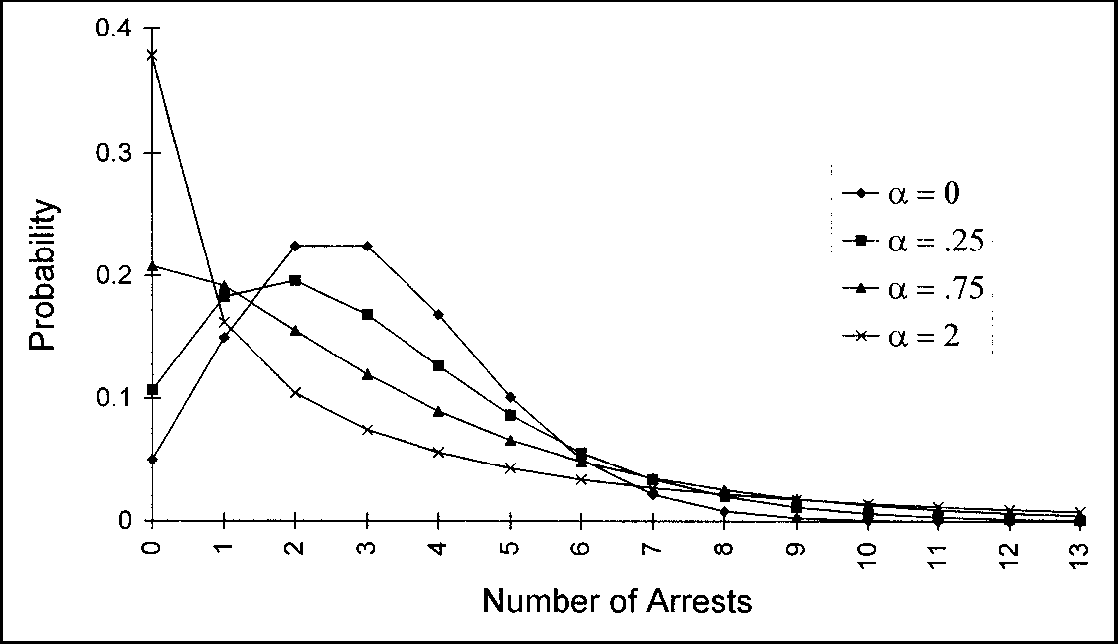
a). A study was conducted by epidemiologist to determine the extent to which the occurrence of coronary heart disease (CHD) was associated with initial serum cholesterol level, sex and age. The data collected was analysed using binary logistic regression model and results are as per the table below.

|  |  |  |
| --- | --- | --- |
| Terms fitted in model | Deviance | d.f |
| Constant | 101.21 | 15 |
| Age | 171.05 | 14 |
| Sex | 221.50 | 14 |
| Age + sex | 88.38 | 13 |
| Age + sex +Age\*sex | 77.44 | 12 |
| Age + Sex + Age \* Sex + Cholesterol level | 22.39 | 9 |

Use the result in this table to answer the following questions

1. Formulate the logistic regression model based on the terms fitted (1mk)
2. If the residual deviance was 80.6, which of the terms fitted in the model contributed significantly on the total variability in the CHD? (3marks)
3. What are other important components of logistic output model will be required in order to provide appropriate interpretation of terms fitted in the model? (2 marks)
4. How will you obtain the odds ratio from the logistic regression model with various explanatory variables? (1 mark)
5. List five good properties that makes logistic regression model to be popularly used (5 mks)

b) In a crime study data was collected on number crimes/1000 people. The data was subjected to Poisson and negative binomial distribution. The results showed that data did not conform to Poisson distribution due to overdispersion hence data fitted the negative binomial distribution with mean count of 3, for four levels of residual variance as shown in the figure below.



1. Interpret the implication of each residual variance denoted by alpha values (5mks)
2. What parameters guided the decision to the data analyst that the data failed to fit Poisson distribution? (2 mks)
3. What is the critical property that makes negative binomial to fit in such kind of data sets? (2 mks)

**END**