UNIVERSITI SAINS MALAYSIA

First Semester Examination
Academic Session 2008/2009

November 2008

MST 567 – Categorical Data Analysis
[Analisis Data Berkategori]

Duration : 3 hours
[Masa : 3 jam]

Please check that this examination paper consists of NINE pages of printed material before you begin the examination.

[Sila pastikan bahawa kertas peperiksaan ini mengandungi SEMBILAN muka surat yang bercetak sebelum anda memulakan peperiksaan ini.]

Instructions: Answer **all eight** [8] questions.

*[**Arahan:** Jawab **semua lapan** [8] soalan.]*

1. The three main distributions for categorical data are the Binomial distribution, the Multinomial distribution, and the Poisson distribution.

   (a) Discuss the overdispersion phenomenon either for Binomial or Poisson distribution.
   (b) What is the connection between Poisson and Multinomial distributions?

   [10 marks]

2. Thirty patients with a well defined skin lesion were randomly allocated to two groups. In one group the subjects received a standard drug and in the other group, a new experimental drug. After 4 weeks of follow-up each patient was observed and a determination made as to whether the lesion had disappeared. The results are as follow.

		LESION PRESENT	
		No	Yes
DRUG	Standard	9	6
	Experimental	4	11

   (a) Test the null hypothesis that the true odds ratio is equal to 1.0 against an alternative hypothesis of not equal to 1.0.
   (b) Find a 95% confidence interval for the odds ratio and the relative risk.
   (c) Comment on the confidence interval you calculated from (b).

   [12 marks]

3. The data in the table below for 205 married persons, give the number of cases in which a tall, medium or short man was mated with a tall, medium or short woman.

		Wife			
		Tall	Medium	Short	Totals
	Tall	18	28	14	60
Husband	Medium	20	51	28	99
	Short	12	25	9	46
	Totals	50	104	51	205

   (a) Find a measure of association between the height of husbands and wives.
   (b) Assign ranks of 1, 2, and 3 to the categories tall, medium and short. Then compute the measure of association between the height of husbands and wives based on their ranks.
   (c) By interpreting the quantities computed in parts (a) and (b), discuss how the heights of husbands and wives are associated.

   [14 marks]

1. *Data berkategori dapat diwakili oleh tiga jenis taburan iaitu taburan Binomial, taburan Multinomial dan taburan Poisson.*

   (a) Bincangkan keadaan lebihan serakan (overdispersion) yang berlaku sama ada pada taburan Binomial atau taburan Poisson.
   (b) Apakah perhubungan antara taburan Poisson dan taburan Multinomial?

   [10 markah]

2. *Tiga puluh pesakit yang mengalami kecederaan kulit dibahagikan secara rawak kepada 2 kumpulan. Kumpulan pertama adalah mereka yang menggunakan ubat biasa manakala kumpulan kedua menggunakan ubat ujikaji. Selepas 4 minggu setiap pesakit diperhatikan sama ada kecederaan pada kulit sudah sembuh. Keputusan yang diperhatikan adalah seperti berikut:*

		KECEDERAAN	
		Tidak	*Ya*
UBAT	*Biasa*	*9*	*6*
	Ujikaji	*4*	*11*

   (a) Uji hipotesis nul bahawa nilai nisbah kemungkinan sebenar bersamaan 1 terhadap hipotesis alternatif bahawa nilai nisbah kemungkinan tidak bersamaan 1.
   (b) Cari selang keyakinan 95% bagi nisbah kemungkinan dan risiko relatif.
   (c) Komen tentang selang keyakinan yang didapati pada (b).

   [12 markah]

3. *Data pada jadual di bawah untuk 205 pasangan berkahwin memberi maklumat mengenai bilangan kes di mana lelaki tinggi, sederhana atau pendek berkahwin dengan wanita tinggi, sederhana atau pendek.*

		Isteri			
		Tinggi	*Sederhana*	*Pendek*	*Jumlah*
	Tinggi	*18*	*28*	*14*	*60*
Suami	*Sederhana*	*20*	*51*	*28*	*99*
	Pendek	*12*	*25*	*9*	*46*
	Jumlah	*50*	*104*	*51*	*205*

   (a) Cari ukuran perkaitan antara ketinggian suami dan isteri.
   (b) Berikan pangkat 1,2, dan 3 bagi kategori Tinggi, Sederhana dan Pendek. Kemudian dapatkan semula ukuran perkaitan antara ketinggian suami dan isteri berdasarkan pangkat.
   (c) Berdasarkan ukuran-ukuran perkaitan pada (a) dan (b) bincangkan perkaitan antara ketinggian suami dan isteri

   [14 markah]

4. A sample of women suffering from excessive menstrual bleeding has been taking an analgesic designed to diminish the effects. A new analgesic is claimed to provide greater relief. After trying the new analgesic, 40 women reported greater relief with the standard analgesic, and 60 reported greater relief with the new one.

   (a) Test the hypothesis that the probability of greater relief with the standard analgesic is the same as the probability of greater relief with the new analgesic. Use $\alpha = 0.05$.

   (b) Construct and interpret the 95% Wald and score confidence intervals for the probability of greater relief with the new analgesic.

   (c) The researchers wanted a sufficiently large sample to be able to estimate the probability of preferring the new analgesic to within 0.08, with confidence 0.95. If the true probability is 0.75, how large a sample is needed to achieve this accuracy?

[14 marks]

5. Consider N independent binary random variables $Y_1,\ldots,Y_N$ with $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$. The probability function of Y_i can be written as

$$\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where $y_i = 0 \; or \; 1$.

(a) Show that this probability function belongs to the exponential family of distributions.

(b) Show that the natural parameter is $\log\left(\frac{\pi_i}{1-\pi_i}\right)$.

(c) Show that $E(Y_i) = \pi_i$.

(d) If the link function is $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = x^T\beta$ show that this is equivalent to modeling the probability π as $\pi = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$.

(e) In the particular case where $x^T\beta = \beta_1 + \beta_2 x$, show that $\pi = \frac{e^{\beta_1+\beta_2 x}}{1+e^{\beta_1+\beta_2 x}}$.

[10 marks]

6. (a) Describe the purpose of the link function of a Generalized Linear Model (GLM). What is the identity link? Explain why it is not often used with binomial or Poisson responses.

   (b) Let $Y_1,\ldots,Y_N$ be independent random variables with

$$E(Y_i) = \mu_i = \beta_0 + \log(\beta_1 + \beta_2 x_i); \quad Y_i \sim N(\mu, \sigma^2)$$

for all $i = 1,2,\ldots,N$. Is this a generalized linear model? Give reasons for your answers

[10 marks]

4. *Suatu sampel wanita yang menderita akibat pendarahan berlebihan semasa kedatangan haid telah mengambil analgesik untuk mengurangkan kesannya. Suatu analgesik yang baru didakwa dapat memberi kelegaan yang lebih baik. Setelah mencuba analgesik baru tersebut, 40 orang wanita melaporkan analgesik yang biasa digunakan adalah lebih baik, dan 60 orang melaporkan analgesik yang baru lebih memberi kelegaan.*

   (a) *Uji hipotesis bahawa kebarangkalian kelegaan dengan analgesik biasa adalah sama seperti kebarangkalian daripada analgesik yang baru. Guna* $\alpha = 0.05$.

   (b) *Bina dan tafsir selang keyakinan Wald dan Score 95% bagi kebarangkalian lebih kelegaan dengan analgesik baru.*

   (c) *Para penyelidik menghendaki satu sampel besar yang mampu untuk menganggar kebarangkalian lebih menyukai analgesik baru dalam lingkungan 0.08, dengan aras keyakinan 0.95. Jika kebarangkalian sebenar adalah 0.75, berapa besarkah sampel yang diperlukan?*

   [14 markah]

5. *Pertimbangkan N pembolehubah biner tak bersandar* $Y_1,\ldots,Y_N$ *dengan* $P(Y_i = 1) = \pi_i$ *dan* $P(Y_i = 0) = 1 - \pi_i$. *Fungsi kebarangkalian untuk* Y_i *dapat ditulis seperti berikut*

$$\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

   di mana $y_i = 0$ *atau* 1.

   (a) *Tunjukkan fungsi kebarangkalian ini adalah dari keluarga taburan eksponen.*

   (b) *Tunjukkan parameter semulajadi adalah* $\log\left(\frac{\pi_i}{1-\pi_i}\right)$.

   (c) *Tunjukkan* $E(Y_i) = \pi_i$

   (d) *Jika fungsi hubungan adalah* $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = x^T\beta$ *tunjukkan ianya setara dengan pemodelan kebarangkalian* π *sebagai* $\pi = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$.

   (e) *Bagi suatu kes tertentu di mana* $x^T\beta = \beta_1 + \beta_2 x$, *tunjukkan* $\pi = \frac{e^{\beta_1+\beta_2 x}}{1+e^{\beta_1+\beta_2 x}}$

   [10 markah]

6. (a) *Huraikan tujuan fungsi hubungan satu Model Linear Teritlak(GLM). Apakah hubungan identiti? Jelaskan mengapa hubungan ini tidak selalu digunakan untuk sambutan Binomial dan Poisson.*

   (b) *Andaikan* $Y_1,\ldots,Y_N$ *pembolehubah rawak tak bersandar dengan*

$$E(Y_i) = \mu_i = \beta_0 + \log(\beta_1 + \beta_2 x_i); \quad Y_i \sim N(\mu,\sigma^2)$$

   untuk semua $i = 1,2,\ldots,N$. *Adakah model ini Model Linear Teritlak? Berikan sebab bagi jawapan anda.*

   [10 markah]

7. A study in Florida stated that the death penalty was given in 19 out of 151 cases in which a white killed a white, in 0 out of 9 cases in which a white killed a black, in 11 out of 63 cases in which a black killed a white, and in 6 out of 103 cases in which a black killed a black. The table below shows results of fitting a logit model for death penalty as the response (1 = yes), with defendant's race (1 = white) and victims' race (1 = white) as indicator predictors.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	1	0.3798
Pearson Chi-Square	1	0.1978
Log Likelihood		-209.4783

Analysis Of Parameter Estimates

		Standard	Likelihood Ratio		Chi-
Parameter	Estimate	Error	95% Conf Limit		Square
Intercept	-3.5961	0.5069	-4.7754	-2.7349	50.33
def	-0.8678	0.3671	-1.5633	-0.1140	5.59
vic	2.4044	0.6006	1.3068	3.7175	16.03

LR Statistics

Source	DF	Chi-Square	Pr > ChiSq
def	1	5.01	0.0251
vic	1	20.35	<.0001

(a) Interpret the parameter estimates. Which group is most likely to have the "yes" response? Find the estimated probability in that case.

(b) Interpret the 95% confidence intervals for conditional odds ratios.

(c) Test the effect of defendant's race, controlling for victims' race, using a (i). Wald test, and (ii) likelihood-ratio test. Interpret your answer.

(d) Test the goodness of fit. Model and interpret your answer.

[15 marks]

7. *Suatu kajian di Florida menyatakan bahawa hukuman mati telah diberikan kepada 19 daripada 151 kes di mana kaum kulit putih membunuh kaum kulit putih yang lain, 0 daripada 9 kes di mana kaum kulit putih membunuh kaum kulit hitam, 11 daripada 63 kes di mana kaum kulit hitam membunuh kaum kulit putih, dan 6 daripada 103 kes di mana kaum kulit hitam membunuh kaum kulit hitam yang lain. Jadual di bawah menunjukkan keputusan penyuian model logit untuk hukuman mati sebagai sambutan (1 = Ya), dengan bangsa defendan (1 = kaum putih) dan bangsa mangsa (1 = kaum putih) sebagai peramal petunjuk.*

Criteria For Assessing Goodness Of Fit

Criterion	*DF*	*Value*
Deviance	*1*	*0.3798*
Pearson Chi-Square	*1*	*0.1978*
Log Likelihood		*-209.4783*

Analysis Of Parameter Estimates

Parameter	*Estimate*	*Standard Error*	*Likelihood Ratio 95% Conf Limit*		*Chi-Square*
Intercept	*-3.5961*	*0.5069*	*-4.7754*	*-2.7349*	*50.33*
def	*-0.8678*	*0.3671*	*-1.5633*	*-0.1140*	*5.59*
vic	*2.4044*	*0.6006*	*1.3068*	*3.7175*	*16.03*

LR Statistics

Source	*DF*	*Chi-Square*	*Pr > ChiSq*
def	*1*	*5.01*	*0.0251*
vic	*1*	*20.35*	*<.0001*

(a) *Tafsir anggaran-anggaran parameter. Kumpulan manakah yang paling berkemungkinan untuk memiliki sambutan "Ya"? Anggarkan kebarangkalian untuk kes tersebut.*

(b) *Tafsir selang keyakinan 95% bagi nisbah kemungkinan bersyarat.*

(c) *Uji kesan bangsa defendan, dengan mengawal bangsa mangsa, menggunakan satu (i). Ujian Wald, dan (ii) Ujian nisbah kebolehjadian Tafsirkan jawapan anda.*

(d) *Uji kebagusan penyuaian model dan tafsirkan jawapan anda.*

[15 markah]

8. Table 1 below is from a General Social Survey. White subjects in the sample were asked: (*B*) Do you favor busing Negro/Black and White school children from one school district to another?, (*P*) If your party nominated a Negro/Black for President, would you vote for him if he were qualified for the job?, (*D*) During the last few years, has anyone in your family brought a friend who was a Negro/Black home for dinner? The response scale for each item was (1 =Yes, 2 = No or Do not know). Table 2 shows output from fitting the model (*BD, BP, DP*). Estimates equal 0 at the second category for any variable.

Table 1

		Home (D)	
President (P)	Busing (B)	1	2
1	1	41	65
	2	72	175
2	1	2	9
	2	4	55

Table 2

```
          Criteria For Assessing Goodness Of Fit

                Criterion              DF       Value

                Deviance                1       0.4794
                Pearson Chi-Square      1       0.5196

          Analysis Of Parameter Estimates

      Parameter          DF          Estimate               Std Error

      Intercept           1           3.9950                 0.1346
      president           1           1.1736                 0.1536
      busing              1          -1.7257                 0.3300
      home                1          -2.4533                 0.4306
      president*busing    1           0.7211                 0.3539
      president*home      1           1.5520                 0.4436
      busing*home         1            0.4672                0.2371

                LR Statistics

Source              DF          Chi-Square          Pr > ChiSq

president*busing    1            4.64               0.0313
president*home      1           17.18               <.0001
busing*home         1           3.83                0.0503
```

(a) Analyze the goodness of fit model. Interpret your answer.
(b) Estimate the conditional odds ratios for each pair of variables. Interpret your answer.
(c) Show all steps of the likelihood-ratio test for the BP association, including explaining which loglinear model holds under the null hypothesis. Interpret your answer.
(d) Construct a 95% confidence interval for the BP conditional odds ratio. Interpret your answer.

[15 marks]

8. *Jadual 1 di bawah adalah daripada suatu Soal Selidik Sosial Umum. Perkara-perkara berikut telah di tanya kepąda suatu sampel kaum kulit putih: (B) Adakah anda menyokong kanak-kanak kulit hitam dan kulit putih menaiki bas yang sama ke sekolah dari suatu daerah ke daerah yang lain?, (P) Jika parti anda mencadangkan orang kulit hitam (Negro) untuk Presiden, adakah anda akan mengundi beliau jika beliau layak?, (D) Sepanjang beberapa tahun terakhir, adakah sesiapa dalam keluarga anda membawa balik seorang kawan berkulit hitam untuk makan malam? Skala jawapan bagi setiap soalan adalah (1 =Ya, 2 = Tidak atau Tidak tahu). Jadual 2 menunjukkan keputusan penyuaian model (BD, BP, DP). Anggaran adalah 0 pada kategori kedua setiap pemboleh ubah.*

Jadual 1

		Rumah (D)	
Presiden(P)	*Menaiki bas sama (B)*	*1*	*2*
1	*1*	*41*	*65*
	2	*72*	*175*
2	*1*	*2*	*9*
	2	*4*	*55*

Jadual 2

```
            Criteria For Assessing Goodness Of Fit

                 Criterion              DF        Value

                 Deviance                1       0.4794
                 Pearson Chi-Square      1       0.5196

            Analysis Of Parameter Estimates

   Parameter           DF         Estimate              Std Error

   Intercept            1          3.9950                0.1346
   president            1          1.1736                0.1536
   busing               1         -1.7257                0.3300
   home                 1         -2.4533                0.4306
   president*busing     1          0.7211                0.3539
   president*home       1          1.5520                0.4436
   busing*home          1           0.4672               0.2371

                 LR Statistics

Source            DF          Chi-Square            Pr > ChiSq

president*busing  1            4.64                 0.0313
president*home    1           17.18                 <.0001
busing*home       1            3.83                 0.0503
```

(a) *Analisis dan tafsirkan kebagusan penyuaian model.*

(b) *Anggarkan dan tafsirkan nisbah kemungkinan bersyarat untuk setiap pemboleh ubah.*

(c) *Tunjukkan setiap langkah ujian nisbah kebolehjadian untuk perkaitan antara BP termasuk menerangkan model log-linear yang manakah merujuk kepada hipotesis nul. Juga tafsirkan jawapan anda.*

(d) *Bina dan tafsirkan selang keyakinan 95% untuk nisbah kemungkinan bersyarat BP.*

[15 markah]

- ooo O ooo -