UNIVERSITY OF CAMBRIDGE

M. P*HIL. IN* STATISTICAL SCIENCE

Friday 3 June, 2005 1:30 to 4:30

APPLIED STATISTICS

*Attempt **FOUR** questions.*

*There are **FIVE** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 The data set given below was taken from The Independent, November 21, 2001, where it appeared under the headline 'Supermarkets to defy bar on cheap designer goods'.

```
How prices compare: prices given in UK pounds.
Item                         UK   Sweden  France  Germany  US
Levi 501 jeans              46.16  47.63  42.11  46.06    27.10
Dockers K1 khakis           58.00  54.08  47.22  46.20    32.22
Timberland women's boots   111.00  104.12 89.43  93.36    75.42
DieselKultar men's jeans    60.00   43.35 43.50  44.48    NA
Timberland cargo pants      53.33   48.58 43.54  58.66    31.70
Gap men's sweater           34.50   NA    26.93  27.26    28.76
Ralph Lauren polo shirt     49.99   42.04 36.41  40.26    32.48
H&M cardigan                19.99   17.31 18.17  15.28    NA
```

Discuss carefully the (slightly edited) *R* analysis given below. This reads the data from the file "dprices". You should interpret the commands and the corresponding results.

What would you expect as the result of the final command?

```
> p = scan("dprices")
Read 40 items
> summary(p)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 15.28   32.22   43.54   46.94   53.33  111.00    3.00
> item = scan(,"")
 1: jeans khakis boots Djeans cpants sweater shirt cardigan
 9:
 Read 8 items
> country = scan(,"")
 1: UK Sweden France Germany US
 6:
 Read 5 items
> Item = gl(8,5,length=40, labels=item)
> Country = gl(5,1, length=40, labels= country)
> plot(Country,p)
> plot(Item,p)
> first.lm = lm(p ~ Country+ Item, na.action=na.omit)
> anova(first.lm)
Analysis of Variance Table
```

```
Response: p
          Df  Sum Sq Mean Sq F value    Pr(>F)
Country    4  1115.6   278.9  10.573 3.732e-05 ***
Item       7 16910.2  2415.7  91.583 3.188e-16 ***
Residuals 25   659.4    26.4

> anova(lm(p ~ Item + Country,  na.action=na.omit))
Analysis of Variance Table

Response: p
          Df  Sum Sq Mean Sq F value    Pr(>F)
Item       7 16409.0  2344.1  88.868 4.566e-16 ***
Country    4  1616.7   404.2  15.323 1.859e-06 ***
Residuals 25   659.4    26.4

> summary(first.lm)

   Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
  (Intercept)      50.799      2.826  17.974 8.27e-16 ***
  CountrySweden    -5.224      2.680  -1.949  0.06260 .
  CountryFrance   -10.708      2.568  -4.170  0.00032 ***
  CountryGermany   -7.676      2.568  -2.989  0.00620 **
  CountryUS       -21.328      2.824  -7.554 6.59e-08 ***
  Itemkhakis        5.732      3.248   1.765  0.08984 .
  Itemboots        52.854      3.248  16.272 8.23e-15 ***
  ItemDjeans        2.935      3.477   0.844  0.40661
  Itemcpants        5.350      3.248   1.647  0.11206
  Itemsweater     -11.509      3.473  -3.314  0.00281 **
  Itemshirt        -1.576      3.248  -0.485  0.63177
  Itemcardigan    -27.210      3.477  -7.825 3.51e-08 ***

  Residual standard error: 5.136 on 25 degrees of freedom
  Multiple R-Squared: 0.9647,     Adjusted R-squared: 0.9492
  F-statistic: 62.12 on 11 and 25 DF,  p-value: 2.4e-15

>summary(lm(p~ Item*Country,  na.action=na.omit)) #final command
```

2 Suppose that $y_1, \ldots, y_n$ are independent Poisson observations, with $\mathbb{E}(y_i) = \mu_i$, and $\log \mu_i = \mu + \beta x_i$, for $1 \leqslant i \leqslant n$. Find equations for $(\hat{\mu}, \hat{\beta})$, the maximum likelihood estimators of (μ, β), and hence show that for large n,

$$\text{var}(\hat{\beta}) \simeq \sum \hat{\mu}_i / \left(\sum \hat{\mu}_i \sum x_i^2 \hat{\mu}_i - (\sum x_i \hat{\mu}_i)^2 \right) \ ,$$

where $\hat{\mu}_i = \mu_i(\hat{\mu}, \hat{\beta})$. If a practical problem showed you that

(i) the deviance for fitting the above model was 27.2, with 29 df

(ii) $\hat{\beta} = 6.73, \ se(\hat{\beta}) = 8.04$

what would you conclude?

3 (i) Assume that the n-dimensional observation vector Y may be written

$$\Omega : Y = X\beta + \epsilon$$

where X is a given $n \times p$ matrix of rank p, β is an unknown vector, and

$$\epsilon \sim N_n(0, \sigma^2 I).$$

Let $Q(\beta) = (Y - X\beta)^T (Y - X\beta)$. Show that $Q(\beta)$ is a convex function of β, and find $\hat{\beta}$, the least-squares estimator of β. Show also that

$$Q(\hat{\beta}) = Y^T (I - H) Y$$

where H is a matrix that you should define.

(ii) Let $\hat{\epsilon} = Y - X\hat{\beta}$. Find the distribution of $\hat{\epsilon}$, and discuss how this may be used to perform diagnostic checks of Ω.

(iii) Suppose that your data actually corresponded to the model

$$Y_i \sim N(\mu_i, \sigma_i^2), \ 1 \leqslant i \leqslant n, \quad \text{with } \sigma_i^2 \propto \mu_i^2 \ .$$

How would your diagnostic checks detect this, and what transformation of Y_i would be appropriate?

4 The British Medical Journal, December 1, 2004 published "Prospective cohort study of cannabis use, predisposition for psychosis, and psychotic symptoms in young people" by C. Henquet and others. This included the following table of data, (slightly simplified here).

	Cannabis use at baseline	Number with psychosis outcome	Number without psychosis outcome	Risk of psychotic systems at follow up
p.no	none	294	1642	15 %
	some	59	216	21 %
p.yes	none	47	133	26 %
	some	23	22	51 %

Here the first 2 rows of the table, "p.no" correspond to those with *no* predisposition for psychosis at baseline, and the second 2 rows of the table "p.yes", correspond to those *with* predisposition for psychosis at baseline.

Discuss carefully the *R* output given below. (This output has been slightly edited, in the interests of simplification.)

```
> cannabis = read.table("cannabis", header=T)
> cannabis
  c.use    with without predisposition
  1  none  294    1642             no
  2  some   59     216             no
  3  none   47     133            yes
  4  some   23      22            yes
>
> chisq.test(rbind(c(341,1775),c(82,238)))

Pearson's Chi-squared test

data:  rbind(c(341, 1775), c(82, 238))
X-squared = 16.8618, df = 1, p-value = 4.02e-05

> attach(cannabis) ; tot = with + without
> summary(glm(with/tot ~ c.use + predisposition, binomial, weights=tot))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.73881    0.06284 -27.673  < 2e-16 ***
c.usesome          0.53847    0.14257   3.777 0.000159 ***
predispositionyes  0.82824    0.15632   5.298 1.17e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Residual deviance:  3.0733  on 1  degrees of freedom
AIC: 31.766

Number of Fisher Scoring iterations: 3


> summary(glm(with/tot ~ c.use *predisposition, binomial, weights=tot))
```

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.72009    0.06333 -27.162  < 2e-16 ***
c.usesome                    0.42235    0.15997   2.640 0.008285 **
predispositionyes            0.67989    0.18112   3.754 0.000174 ***
c.usesome:predispositionyes  0.66230    0.37857   1.749 0.080209 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Residual deviance: 2.8377e-13  on 0  degrees of freedom
AIC: 30.692

Number of Fisher Scoring iterations: 3
```

5 The table below summarises the results from an hypothetical randomised controlled trial to determine the cost-effectiveness of a new drug regime versus the standard regime, which is already in use in the UK National Health Service (NHS). The 2 regimes are designed to prolong the survival time of patients with small cell lung cancer. If the new drug regime is found to offer good value for money by the National Institute of Clinical Excellence (NICE), then it may be approved for reimbursement under the NHS.

(i) Estimate the incremental cost-effectiveness ratio (ICER) from the table (providing the appropriate units of the ICER).

(ii) Show how Fieller's method may be used to calculate a 95% confidence interval (CI) for the ICER (stating any assumptions made). Suggest an alternative method of constructing a 95% CI for the ICER.

(iii) What are some of the problems with interpreting cost-effectiveness ratios?

(iv) Describe two other methods used for presenting cost-effectiveness results.

	New Drug Regime ($n = 64$)	Standard Drug Regime ($n = 65$)	Difference in Means
Mean (SE) survival times in days	368 days (se = 12 days)	323 days (se = 9 days)	$\Delta E = 45$ days (se = 15 days)
Mean (SE) costs in £	£2840 (se=£240)	£1040 (se = £180)	$\Delta C =$ £1800 (se=£300)
Correlation between ΔE and ΔC			0.5

END OF PAPER