UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 31 May, 2012 9:00 am to 12:00 pm

PAPER 37

APPLIED STATISTICS

*Attempt no more than **FOUR** questions, with at most **THREE** from Section A.*

*There are **SIX** 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.

SECTION A

1

(a) Write down the form of an ordinary linear model for a $n \times 1$ response vector $\boldsymbol{Y}$, $n \times p$ covariate matrix $\boldsymbol{X}$ of rank p, and $p \times 1$ parameter vector $\boldsymbol{\beta}$. Derive the form of the maximum likelihood estimator for $\boldsymbol{\beta}$, and any other parameters.

(b) Define the residuals, and derive their joint distribution under the model. [*Hint: You may use the fact that if* $\boldsymbol{Z} \sim N(\mathbf{0}, \Sigma)$ *for some vector* $\boldsymbol{Z}$*, then* $A\boldsymbol{Z} \sim N(\mathbf{0}, A\Sigma A^T)$ *for any suitable matrix* A*.*]

A nutritionist measures the weights and heights of $n = 100$ students, $n_1 = 45$ female and $n_2 = 55$ male. Let the weight of student $j = 1, \ldots, n_i$ of sex $i = 1, 2$ be Y_{ij}, and their height be x_{ij}. The nutritionist would like to be able to predict the weight of other students, as closely as possible, from their sex and height.

(c) Explain the meaning and purpose of using the command `I(height-165)` in the `R` output below. Write out the model `lm2` algebraically, and give the estimates and standard errors for each regression parameter; make sure you interpret the results qualitatively in terms of the original data.

(d) Considering all the `R` output, write a concise summary for the nutritionist about your findings. For any hypothesis tests you use, state the hypotheses and null distribution. Comment in particular on the difference between the slope parameters for height in models `lm1` and `lm2`.

(e) The summaries give the maximum of the residuals for each model; by comparing this to the residual standard errors, suggest any comments you might make to the nutritionist about the data and the models.

```
> head(dat)
  height sex weight
1    163   F     62
2    171   F     53
3    167   M     64
4    177   M     58
5    174   M     72
6    159   F     60
> options()$contrasts
        unordered           ordered
"contr.treatment"      "contr.poly"
> lm1 = lm(weight ~ I(height-165), data=dat)
> lm2 = lm(weight ~ sex + I(height-165), data=dat)
> lm3 = lm(weight ~ sex*I(height-165), data=dat)
> summary(lm1)

Residuals:
     Min       1Q   Median       3Q      Max
-12.4082  -3.3276   0.0455   3.7609  26.6071

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      59.8853     0.7418  80.729  < 2e-16
I(height - 165)   0.8769     0.1175   7.463 3.48e-11

Residual standard error: 5.942 on 98 degrees of freedom
Multiple R-squared: 0.3624,Adjusted R-squared: 0.3559
F-statistic:  55.7 on 1 and 98 DF,  p-value: 3.477e-11

> summary(lm2)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2536  -3.4587   0.0257   3.2571  24.7151

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      58.4711     0.8533  68.520  < 2e-16
sexM              4.3294     1.4356   3.016  0.00327
I(height - 165)   0.6211     0.1412   4.397 2.81e-05

Residual standard error: 5.71 on 97 degrees of freedom
Multiple R-squared: 0.4171,Adjusted R-squared: 0.405
F-statistic:  34.7 on 2 and 97 DF,  p-value: 4.294e-12

> summary(lm3)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-11.9844  -3.5162  -0.0463   3.1702  24.5908

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          58.4506     0.8593  68.020  < 2e-16
sexM                  4.6710     1.7504   2.668  0.00895
I(height - 165)       0.6696     0.2000   3.349  0.00116
sexM:I(height - 165) -0.0977     0.2838  -0.344  0.73139

Residual standard error: 5.737 on 96 degrees of freedom
Multiple R-squared: 0.4178,Adjusted R-squared: 0.3996
F-statistic: 22.96 on 3 and 96 DF,  p-value: 2.737e-11
```

2

A farmer is measuring the yields of two varieties of wheat under various conditions. He has 4 fields, each of equal size. In each of the three years 2004, 2005, 2006, he planted each of the two varieties of wheat in two fields, and gave each either a high or low dose of fertiliser. The data are given in the table below.

Variety	**Fertiliser**	**Year**		
		2004	2005	2006
1	Low	5.26	6.93	5.99
	High	8.08	9.15	8.36
2	Low	9.91	11.03	8.19
	High	6.39	8.84	7.45

Explain what is meant by a factor, in the context of linear models. Why might we want to treat the year as a factor when analysing the wheat yields?

Look at the edited `R` output below. Write down the model being fitted in `lmSF.Y` algebraically, remembering to define each symbol you use, and any constraints.

In the output from the `anova()` command below, some of the values have been removed and replaced with `xx`. Rewrite the ANOVA table, filling in the missing entries from the columns `Df` and `Mean Sq`; explain how to calculate the missing entry in the column labelled `F value`, and give a rough approximation to its value. What is the hypothesis test being tested in the row labelled `seed:fert`? State the hypotheses, null distribution, and your conclusion clearly. Based on this ANOVA table only, would you suggest a simpler model to try to fit? Explain your answer.

Now look at the output from the `summary()` command. How would you interpret the results of this model fit, and what would you tell the farmer about the yields from the two varieties of wheat under the different conditions?

```
> seed
 [1] 1 1 1 1 1 1 2 2 2 2 2 2
Levels: 1 2
> fert
 [1] Low  Low  Low  High High High  Low  Low  Low High High High
Levels: Low High
> year
 [1] 2004 2005 2006 2004 2005 2006 2004 2005 2006 2004 2005 2006
Levels: 2004 2005 2006
> yield
 [1]  5.26  6.93  5.99  8.08  9.15  8.36  9.91 11.03  8.19  6.39
 [11] 8.84  7.45

> lmSF.Y = lm(yield ~ seed*fert + year)
> anova(lmSF.Y)
Analysis of Variance Table
```

```
Response: yield
          Df  Sum Sq Mean Sq F value   Pr(>F)
seed      xx  5.3868      xx 11.3748 0.014993
fert      xx  0.0768      xx  0.1622 0.701122
year      xx  6.2883      xx  6.6392 0.030146
seed:fert xx 16.0083      xx      xx 0.001137
Residuals xx  2.8414      xx

> summary(lmSF.Y)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.5050     0.4866  11.313 2.85e-05
seed2            3.6500     0.5619   6.496 0.000633
fertHigh         2.4700     0.5619   4.396 0.004589
year2005         1.5775     0.4866   3.242 0.017646
year2006         0.0875     0.4866   0.180 0.863216
seed2:fertHigh  -4.6200     0.7946  -5.814 0.001137
---

Residual standard error: 0.6882 on 6 degrees of freedom
Multiple R-squared: 0.9071,Adjusted R-squared: 0.8298
F-statistic: 11.72 on 5 and 6 DF,  p-value: 0.004715
```

3

Adult beetles of the genus *Tribolium* eat the eggs of their own species as well as those of closely related species. An experiment was conducted on adult beetles of three species, denoted by A, B, and C, to determine if any of these species has evolved a preference for eggs of the other species, that is, if any of the three species can recognise and avoid eggs of its own species - clearly an evolutionary advantage!

The experiment was conducted on three different days with a number of repetitions. A number of adult beetles of each species was isolated, and presented with a vial of 150 eggs, 50 of each type. Two days later, the number of eggs remaining of each type was recorded. The data are presented in the table below.

	Adult species					
	A		B		C	
	Eggs	Total	Eggs	Total	Eggs	Total
Day 1	61	86	53	88	44	66
	60	84	36	64	45	61
	44	68	46	84	51	70
Day 2	54	75	73	115	52	79
	68	90	62	100	38	61
	65	88	-	-	30	53
Day 3	77	101	73	117	-	-
	78	108	63	102	-	-
	63	99	80	125	-	-

The column *Eggs* gives the number of eggs of the other species that were eaten over the 2 day period, and the column *Total* gives the total number of eggs eaten. So for species A, *Eggs* gives the number of eggs of species B and C that were eaten. The data show that only 2 experiments were conducted for species B in Day 2, and none for species C in day 3. The `R` output below contains an analysis of these data.

(a) Write down the algebraic form of the model fitted in `eggs.glm1`, defining your notation carefully and giving any constraints explicitly.

(b) In the analysis of deviance table, six entries have been replaced by `xx`. Find these values.

(c) Using deviances, test whether the probability of eating eggs of the other species depends on the day in which the experiment started, controlling for Species. Give the null hypothesis, the test statistic, its null distribution, the result, and your conclusion in words.

(d) Using your preferred model, obtain an estimate of the log-odds of eating eggs of the other species for an adult beetle of Species A. Explain how to obtain an approximate 95% confidence interval for this log-odds, stating any asymptotic distribution results used.

(e) The confidence interval in (d) equals $(0.7586, 1.065)$. What do you conclude about the eating preferences of adult beetles of Species A?

```
> eggs <- c(61,60,44,54,68,65,77,78,63,53,36,46,73,62,73,63,80,44,45,51,
+ 52,38,30)
> total <- c(86,84,68,75,90,88,101,108,99,88,64,84,115,100,
+ 117,102,125,66,61,70,79,61,53)
> prop <- eggs / total
> Species <- as.factor(c(rep("A",times=9),rep("B",times=8),rep("C",times=6)))
> Day <- as.factor(c(rep(1,3), rep(2,3), rep(3,3), rep(1,3), rep(2,2),
+ rep(3,3), rep(1,3), rep(2,3)))
> eggs.glm1 <- glm(prop ~ Day + Species, family=binomial, weights=total)
> summary(eggs.glm1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.87083    0.10612   8.206 2.28e-16
Day2         0.04578    0.11641   0.393    0.694
Day3         0.06941    0.12270   0.566    0.572
SpeciesB    -0.46029    0.10701  -4.301 1.70e-05
SpeciesC    -0.20026    0.13919  -1.439    0.150
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 33.631  on 22  degrees of freedom
Residual deviance: 14.646  on 18  degrees of freedom
> anova(eggs.glm1, test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: prop
Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       xx     xx
Day      xx  xx            20     33.358    0.8723
Species  2   xx            18     xx      8.647e-05
> eggs.glm2 <- glm(prop ~ Species, family=binomial, weights=total)
> summary(eggs.glm2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.91191    0.07824  11.656  < 2e-16
SpeciesB    -0.45905    0.10684  -4.297 1.73e-05
SpeciesC    -0.21877    0.13289  -1.646   0.0997
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 33.631  on 22  degrees of freedom
Residual deviance: 14.987  on 20  degrees of freedom
```

4

Suppose that we have independent Poisson observations $Y_1, \ldots, Y_n$ with

$$\mathbb{P}(Y_i = y) = \frac{e^{-\mu_i}\mu_i^y}{y!}, \quad y = 0, 1, \ldots,$$

and the means μ_i are positive. Assume that we fit the following model

$$\log \mu_i = x_i^T \beta = \sum_{j=1}^{p} x_{ij}\beta_j,$$

where β is a p-dimensional vector of unknown parameters, and x_i is a p-dimensional vector of known covariates for observation i. Let x_i^T denote the transpose of x_i.

(a) Write down the log-likelihood for this model and derive a set of equations that must be satisfied by $\hat{\beta}$, the maximum likelihood estimator of β.

(b) Derive an expression for the deviance of this model. How and when can the deviance be used to test for goodness of fit?

(c) The following data set contains information on 22 rats that received a single dose of the carginogen azoxymethane (AOM), a chemical that induces colon cancer. The rats were sacrificed at three different times, 6, 12 and 18 weeks after injection, and their colons were removed and analysed. The count of aberrant crypt foci observed in the colon of each rat is recorded, where these crypts are distinctly different in size and shape, as well as thickness of lining, from crypts of healthy animals. It is believed that these aberrant crypt foci represent precursor lesions of chemically induced colon cancer. Let *count* denote the number of aberrant crypt foci, and *endtime* denote the time, following injection, when the rat was sacrificed. The `R` output below contains an analysis of these data.

```
> count <- c(1,3,5,1,2,1,1,3,1,2,6,0,0,4,1,10,6,6,7,5,7,6)
> endtime <- c(6,6,6,6,6,6,6,12,12,12,12,12,12,12,12,18,18,18,18,18,
+ 18,18)
> aom.glm1 <- glm(count ~ endtime, family=poisson)
> aom.glm2 <- glm(count ~ endtime + I(endtime^2), family=poisson)
> anova(aom.glm1, aom.glm2, test="Chisq")
Model 1: count ~ endtime
Model 2: count ~ endtime + I(endtime^2)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        20     28.369
2        19     24.515  1   3.8548    0.0496
```

What assumptions are made to justify fitting a Poisson model to this data? Write down the algebraic form of the model fitted in `aom.glm2`. What test of hypothesis is performed by the `anova` command above? Give the null hypothesis, the test statistic, its null distribution, the result, and your conclusion in words. Does it make sense to test for the goodness of fit of model `aom.glm2`?

Explain what is meant by the term over-dispersion. Give one plausible reason why this data might display over-dispersion. We proceed to estimate the dispersion

parameter by the following formula.

$$\hat{\phi} = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \Big/ (n - p),$$

where $Y_i = count_i$ is the response.

```
> phi <- sum((count-acf.glm2$fitted.values)^2/acf.glm2$fitted.values)/19
> phi
[1] 1.254068
```

How would you modify the test of hypothesis above to take account of over-dispersion? Give the test statistic and its null distribution.

SECTION B

5

(a) In survival data analysis, what is

(i) right censoring?

(ii) the relationship between the censoring distribution and the survival time distribution if right censoring is assumed to be *non-informative*?

(iii) implied, at each time point, for those subjects who are censored and those who are still under observation when right censoring is assumed *non-informative* in a study?

(b) Let T be a positive discrete survival time random variable which takes unique ordered failure time points $t_1, t_2, \ldots, t_n$ (i.e. $0 = t_0 < t_1 < \ldots < t_n$). In a study of time to failure of m subjects, interest lies in estimating the survivor function, $S(t) = \mathbb{P}(T > t)$, of T. Denote by d_j the number of subjects that are observed to fail at t_j and n_j the number of subjects who were still at risk immediately prior to t_j. Define $f(t_j)$ and $h(t_j)$ to be the probability mass function and discrete hazard function of failing at t_j respectively. The latter corresponds to the conditional probability of failing at t_j, given still alive beyond the previous failure time point t_{j-1}. Assume that censoring in this study is non-informative.

Derive an expression for the survivor function at time t in terms of only the discrete hazard function. From this, write down the "Kaplan-Meier" estimator of the survivor function.

(c) In a recently completed randomised controlled clinical trial investigating the effect of chemotherapy versus radiotherapy on time to death after surgical removal of a localised cancerous lung tumour in elderly patients, both the time to death from lung cancer and the time to death from other causes since surgery are of interest. In particular, the estimated probabilities of dying from lung cancer and dying from other causes within 5 years of surgery, and the estimated effect of treatment on survival are required. The data collected are in the form (T_i, D_i), where T_i represents either the time to death, if observed, or the time until the end of study if death has been unobserved for the ith patient, and is recorded in the variable `time` within the data-set, `cancer.dat`. D_i takes the value 0, 1 or 2 depending on whether death is unobserved (`status`= 0), death is observed to be due to lung cancer (`status`= 1) or death is observed to be from other causes (`status`= 2) for the ith patient.

Dr PH Cox, the study statistician, is responsible for analysing the data. Initially, Dr Cox decides to describe the data, ignoring treatment group status, by producing two overall "Kaplan-Meier"-type survival curves. The first curve is for time to death from lung cancer, where deaths from other causes are treated as censored data. The second is for the time to death from other causes where now deaths from lung cancer are treated as censored data. The data from subjects who have not died by the end of the study are treated as censored in both situations. Afterwards Dr Cox analyses time to death from lung cancer and

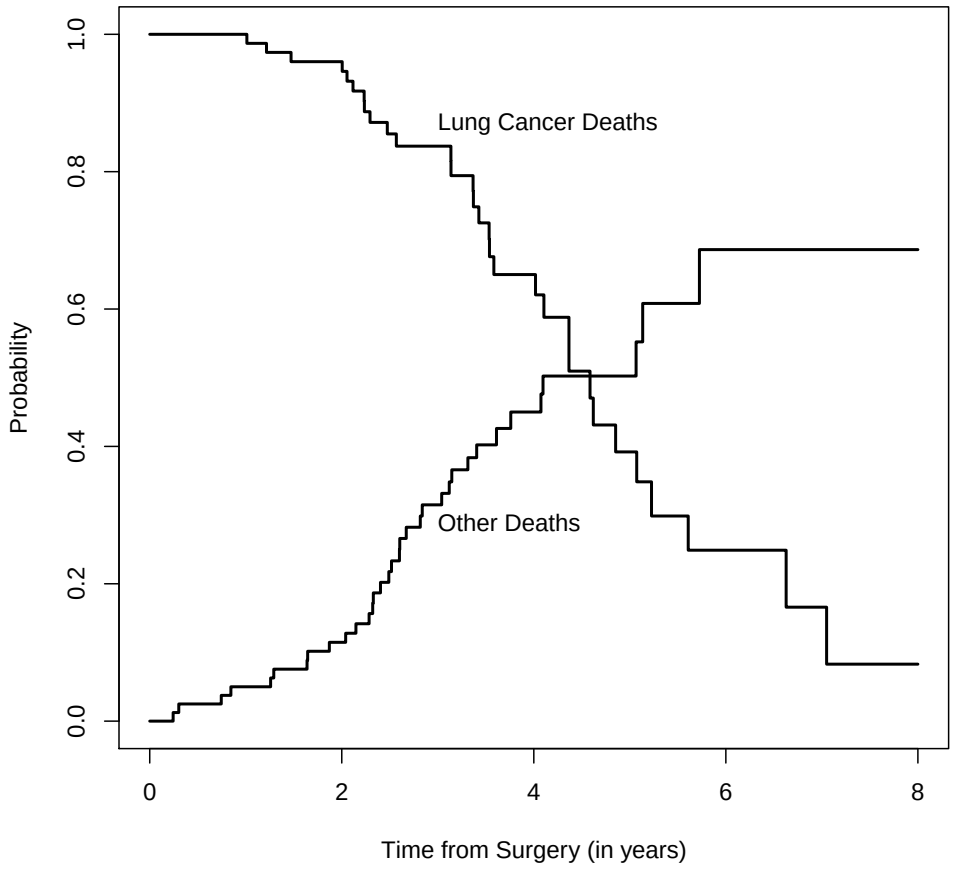
time to death from other causes by fitting two separate proportional hazards models with type of treatment (i.e. chemotherapy versus radiotherapy, where `trt`= 0 corresponds to radiotherapy and `trt`= 1 corresponds to chemotherapy) as the only included explanatory variable.

The plots produced by Dr Cox using R are shown in the accompanying Figure. The corresponding "Kaplan-Meier" tables are provided also. For ease of presentation, the information in the figure for the estimated survival curve for death from other causes is displayed in the form of a cumulative distribution function (i.e. one minus the survivor function), while the information in the estimated survivor function for death from lung cancer is displayed in the standard way as a survival curve. The R code and edited R output of results from the proportional hazards models fitted by Dr Cox also are displayed.

(i) From the provided plots, the details of which are presented in the "Kaplan-Meier" tables, give the estimated overall probabilities of dying from lung cancer and dying from other causes within 5 years of surgery. Are these unbiased estimates for $\mathbb{P}(T \leqslant 5, D = 1)$ and $\mathbb{P}(T \leqslant 5, D = 2)$? Give reasons to justify your answer.

(ii) Give the multi-state diagram representation for the analysis performed by Dr Cox. The transition intensity functions should be included in the diagram, with their mathematical equations describing the estimated relationships between rates of transition between the various states and treatment given. All notation used must be defined.

(iii) Interpret for the study investigators (who may not be statisticians) the treatment effect obtained from the fitted proportional hazards model corresponding to time to death from lung cancer.


Probability
0.0
0.2
0.4
0.6
0.8
1.0
Lung Cancer Deaths
Other Deaths
0
2
4
6
8
Time from Surgery (in years)

```
> lungcancerdeath.km <- survfit(Surv(time,(status==1))~1,data=cancer.dat)
> summary(lungcancerdeath.km)
Call: survfit(formula = Surv(time, (status == 1)) ~ 1, data = cancer.dat)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 1.01     76       1    0.987  0.0131       0.9616        1.000
 1.22     75       1    0.974  0.0184       0.9384        1.000
 1.47     72       1    0.960  0.0225       0.9170        1.000
 2.00     68       1    0.946  0.0263       0.8959        0.999
 2.05     66       1    0.932  0.0295       0.8756        0.991
 2.12     65       1    0.917  0.0324       0.8561        0.983
 2.23     61       1    0.902  0.0351       0.8360        0.974
 2.24     60       1    0.887  0.0376       0.8165        0.964
 2.29     57       1    0.872  0.0401       0.7966        0.954
 2.47     52       1    0.855  0.0427       0.7753        0.943
 2.57     48       1    0.837  0.0453       0.7528        0.931
 3.14     39       1    0.816  0.0490       0.7251        0.918
 3.14     38       1    0.794  0.0522       0.6982        0.903
 3.37     35       1    0.772  0.0554       0.6702        0.888
 3.37     34       1    0.749  0.0583       0.6429        0.872
 3.43     32       1    0.725  0.0610       0.6153        0.855
 3.53     30       1    0.701  0.0635       0.5872        0.838
 3.54     28       1    0.676  0.0660       0.5584        0.819
 3.58     26       1    0.650  0.0684       0.5290        0.799
 4.02     22       1    0.621  0.0714       0.4954        0.778
 4.11     19       1    0.588  0.0747       0.4583        0.754
 4.37     15       1    0.549  0.0794       0.4133        0.729
 4.37     14       1    0.510  0.0828       0.3706        0.701
 4.59     13       1    0.470  0.0852       0.3298        0.671
 4.62     12       1    0.431  0.0867       0.2908        0.639
 4.85     11       1    0.392  0.0872       0.2535        0.606
 5.07      9       1    0.348  0.0877       0.2127        0.571
 5.23      7       1    0.299  0.0882       0.1674        0.533
 5.61      6       1    0.249  0.0864       0.1260        0.491
 6.63      3       1    0.166  0.0889       0.0580        0.474
 7.05      2       1    0.083  0.0736       0.0146        0.472
```

```
> otherdeath.km <- survfit(Surv(time,(status==2))~1,data=cancer.dat)
> summary(otherdeath.km)
Call: survfit(formula = Surv(time, (status == 2)) ~ 1, data = cancer.dat)

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.244     80       1    0.988  0.0124        0.963        1.000
 0.303     79       1    0.975  0.0175        0.941        1.000
 0.745     78       1    0.963  0.0212        0.922        1.000
 0.845     77       1    0.950  0.0244        0.903        0.999
 1.259     74       1    0.937  0.0272        0.885        0.992
 1.292     73       1    0.924  0.0297        0.868        0.984
 1.638     71       1    0.911  0.0320        0.851        0.976
 1.645     70       1    0.898  0.0341        0.834        0.968
 1.871     69       1    0.885  0.0360        0.817        0.959
 2.041     67       1    0.872  0.0378        0.801        0.949
 2.148     63       1    0.858  0.0397        0.784        0.940
 2.285     58       1    0.843  0.0417        0.766        0.929
 2.322     56       1    0.828  0.0435        0.747        0.918
 2.328     55       1    0.813  0.0453        0.729        0.907
 2.403     53       1    0.798  0.0470        0.711        0.896
 2.490     51       1    0.782  0.0486        0.693        0.884
 2.518     50       1    0.767  0.0501        0.675        0.871
 2.601     47       1    0.750  0.0516        0.656        0.859
 2.604     46       1    0.734  0.0530        0.637        0.846
 2.672     45       1    0.718  0.0543        0.619        0.832
 2.818     44       1    0.701  0.0554        0.601        0.819
 2.838     43       1    0.685  0.0565        0.583        0.805
 3.041     41       1    0.668  0.0575        0.565        0.791
 3.119     40       1    0.652  0.0585        0.547        0.777
 3.146     37       1    0.634  0.0595        0.528        0.762
 3.313     36       1    0.616  0.0604        0.509        0.747
 3.405     33       1    0.598  0.0614        0.489        0.731
 3.611     25       1    0.574  0.0634        0.462        0.713
 3.760     24       1    0.550  0.0651        0.436        0.694
 4.074     21       1    0.524  0.0671        0.408        0.673
 4.096     20       1    0.498  0.0686        0.380        0.652
 5.065     10       1    0.448  0.0777        0.319        0.629
 5.134      8       1    0.392  0.0858        0.255        0.602
 5.725      5       1    0.313  0.0981        0.170        0.579
```

```
> lungcancerdth.cox <- coxph(Surv(time,(status==1))~trt,data=cancer.dat)
> summary(lungcancerdth.cox)
Call:
coxph(formula = Surv(time, (status == 1)) ~ trt, data = cancer.dat)

  n= 80, number of events= 31

      coef exp(coef) se(coef)     z Pr(>|z|)
trt 1.3113    3.7111   0.4346 3.018  0.00255 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    exp(coef) exp(-coef) lower .95 upper .95
trt     3.711     0.2695     1.583     8.697


> otherdeath.cox <- coxph(Surv(time,(status==2))~trt,data=cancer.dat)
> summary(otherdeath.cox)
Call:
coxph(formula = Surv(time, (status == 2)) ~ trt, data = cancer.dat)

  n= 80, number of events= 34

      coef exp(coef) se(coef)     z Pr(>|z|)
trt 0.5903    1.8045   0.3640 1.621    0.105

    exp(coef) exp(-coef) lower .95 upper .95
trt     1.805     0.5542    0.8841     3.683
```

6

A behavioural researcher approaches two statisticians (Dr Quasi and Dr Pascal) with data collected from a one academic year follow-up study of m Part IIB and Part III Mathematical Tripos students. The researcher has collected data on the number of times each student in the study has been to any Night Club in Cambridge during each of the three terms (Michaelmas, Lent and Easter), and is interested in modelling these count profiles over time and determining what variables can affect the number of visits in a term.

The researcher has brought the Night Club data to the two statisticians in the form of a vector of counts $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$ for the ith subject chronologically ordered over the three terms in the academic year. Additional information on the student is recorded in a baseline covariate vector $\mathbf{x}_i$, which includes details on the distance of the student's college to the nearest Night Club, whether the student is reading for Part IIB or Part III of the Mathematical Tripos and the gender of the student. Furthermore, the term effect is represented by the factor variable, α_j ($j = 1, 2, 3$ with $1, 2, 3$ corresponding to Michaelmas, Lent, Easter respectively). The researcher believes that there may be some students who over their entire time at the University will never be inclined towards attending a Night Club during term time.

Both Drs Quasi and Pascal recognise that there will be correlation between the components of $\mathbf{Y}_i$. Moreover they realise that they need to incorporate into their models the researcher's belief that there is a proportion of students who will never be inclined towards attending a Night Club during term time.

Dr Quasi decides to model the data as follows. He assumes that

$$\begin{aligned}
\log \mathbb{E}(Y_{ij}|\mathbf{x}_i, \alpha_j, \beta_0, \beta^T, \pi) &= \log(\mu_{ij}(1-\pi)) = \beta_0 + \beta^T \mathbf{x}_i + \alpha_j + \log(1-\pi) \\
Var(Y_{ij}|\mathbf{x}_i, \alpha_j, \beta_0, \beta^T, \pi) &= (1-\pi)\mu_{ij}(1 + (\pi+\tau)\mu_{ij}) \\
Cov(Y_{ij}, Y_{ik}|\mathbf{x}_i, \alpha_j, \alpha_k, \beta_0, \beta^T, \pi) &= \mu_{ij}\mu_{ik}\tau \quad (j \neq k),
\end{aligned}$$

where $\pi \in [0, 1]$ and $\tau \geqslant 0$ are parameters that accommodate the possibility a student will never be inclined to attend a Night Club during term time and potential unexplained heterogeneity respectively.

However, Dr Pascal decides to adopt the following alternative strategy. She assumes that conditional on a random effect b_i, the counts $\{Y_{ij} : j = 1, 2, 3\}$ on the ith student are independent Poisson random variables with

$$\begin{aligned}
\mathbb{E}(Y_{ij}|b_i; \mathbf{x}_i, \alpha_j, \beta_0, \beta^T) &= \eta_{ij} = b_i \exp(\beta_0 + \beta^T \mathbf{x}_i + \alpha_j) \\
Var(Y_{ij}|b_i; \mathbf{x}_i, \alpha_j, \beta_0, \beta^T) &= \eta_{ij} \\
Cov(Y_{ij}, Y_{ik}|b_i; \mathbf{x}_i, \alpha_j, \alpha_k, \beta_0, \beta^T) &= 0 \quad (j \neq k).
\end{aligned}$$

She assumes that the b_i's are independently and identically distributed random variables and that b_i is from a mixture distribution with probability density function

$$f_{b_i}(u) = \begin{cases} \pi & \text{if } u = 0 \\ (1-\pi)\frac{(1/\tau)^{1/\tau}}{\Gamma(1/\tau)} u^{(1/\tau)-1} \exp(-u/\tau) & \text{if } u > 0 \end{cases}$$

That is, b_i is from a mixture distribution with a point mass at zero and for values greater than zero, a Gamma$(1/\tau, 1/\tau)$ distribution with mean and variance 1 and τ respectively. The mixing probabilities are π and $1-\pi$ respectively.

(a) What are the differences between the two approaches?

(b) What estimating approaches would Drs Quasi and Pascal base their inferences upon?

(c) How would Drs Quasi and Pascal *correctly* interpret their respective $\{\alpha_j\}$ parameters for the behavioural researcher? They have both assumed that $\alpha_1 = 0$ in their models to avoid non-identifiability.

(d) Work out the unconditional distribution for Y_{ij} and the first two moments of $\mathbf{Y}_i$ from Dr Pascal's model. If Dr Pascal's model was the "correct" model for analysing these Night Club data, would Dr Quasi be consistently *estimating* what he *thinks* he is estimating? Why?

(e) Dr Pascal has fitted her model above to the data collected and obtained a log-likelihood value of -1201.56. She then refitted her model further assuming $\pi = 0$, and obtained a log-likelihood value of -1202.92. Test the null hypothesis, $H_0 : \pi = 0$, against the alternative, $H_1 : \pi > 0$, showing all your calculations. Use a 5%-significance level for your hypothesis test.

You may find the following useful.

If X is from a Gamma(a, b) distribution, then the probability density function of X is given by

$$f_X(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

For random variables X_1, X_2, X_3 with $\mathbb{E}(X_i^2) < \infty$ $(i = 1, 2, 3)$, we have

$$Cov(X_1, X_2) = \mathbb{E}[Cov(X_1, X_2|X_3)] + Cov[\mathbb{E}(X_1|X_3), \mathbb{E}(X_2|X_3)]$$

The 90th and 95th percentiles for a Chi-squared distribution with $1, 2$ and 3 degrees of freedom are given by

$$\chi^2_{0.90}(1) = 2.71 \text{ and } \chi^2_{0.95}(1) = 3.84$$
$$\chi^2_{0.90}(2) = 4.61 \text{ and } \chi^2_{0.95}(2) = 5.99$$
$$\chi^2_{0.90}(3) = 6.25 \text{ and } \chi^2_{0.95}(3) = 7.81$$

END OF PAPER