MATHEMATICAL TRIPOS Part III

Tuesday 3 June 2008 1.30 to 4.30

PAPER 41

APPLIED STATISTICS

Attempt no more than **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 Let $Y = X\beta + \epsilon$ where $Y^T = (Y_1, \ldots, Y_n)$, $\beta^T = (\beta_1, \ldots, \beta_p)$, X is a known $n \times p$ matrix with rank p ($< n$), and $\epsilon^T = (\epsilon_1, \ldots, \epsilon_n)$, where $\epsilon_1, \ldots, \epsilon_n$ are independent normal random variables with mean zero and variance σ^2. Find the least squares estimator $\hat{\beta}$ of β and find its distribution. Define the residual sum of squares RSS and write down an estimator for σ^2. Explain how to test $H_0 : \beta_1 = 0$ in the above model.

In the edited R output below, `NO2` contains observations of the nitrogen dioxide concentration in a particular location for 25 days, and, for the same 25 days, `wind` contains the average windspeed (in miles per hour), `maxtemp` contains the maximum temperature (in degrees Fahrenheit), and `insol` contains the insolation (a measure of solar radiation energy, in langleys per day). Write down the model fitted in `model1.lm` and interpret in detail the output to `summary(model1.lm)`.

```
> model1.lm <- lm(NO2 ~ wind + maxtemp + insol)
> summary(model1.lm)

Call:
lm(formula = NO2 ~ wind + maxtemp + insol)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3052 -1.1710 -0.4990  0.9823  3.4033

Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)

(Intercept)  3.784916   7.979647    0.474   0.6402
wind        -0.527410   0.224904   -2.345   0.0289
maxtemp      0.124991   0.075173    1.663   0.1112
insol       -0.005259   0.006637   -0.792   0.4370

Residual standard error: 1.844   on 21 degrees of freedom
Multiple R-Squared:      0.6533, Adjusted R-squared: 0.6037
```

In the edited output below, explain the output to the `stepAIC` directive. Comment briefly on the output to `boxcox(model1.lm)` which is shown in Figure 1.

```
> library(MASS)
> stepAIC(model1.lm,
           scope = list(upper = ~ wind + maxtemp + insol, lower=~1), test="F")

Start:  AIC = 34.25
 NO2 ~ wind + maxtemp + insol

          Df Sum of Sq    RSS    AIC F Value   Pr(F)
- insol    1     2.136 73.564 32.982   0.628 0.43700
<none>                 71.428 34.246
- maxtemp  1     9.403 80.832 35.337   2.765 0.11123
- wind     1    18.705 90.133 38.060   5.499 0.02893
```

```
Step:   AIC = 32.98

 NO2 ~ wind + maxtemp

          Df Sum of Sq    RSS    AIC  F Value   Pr(F)
<none>                  73.564 32.982
- maxtemp  1     7.434 80.998 33.389   2.223  0.15016
+ insol    1     2.136 71.428 34.246   0.628  0.43700
- wind     1    23.684 97.248 37.960   7.083  0.01426

Call: lm(formula = NO2 ~ wind + maxtemp)

Coefficients:
(Intercept)         wind      maxtemp
     4.4368      -0.5734       0.1040
```

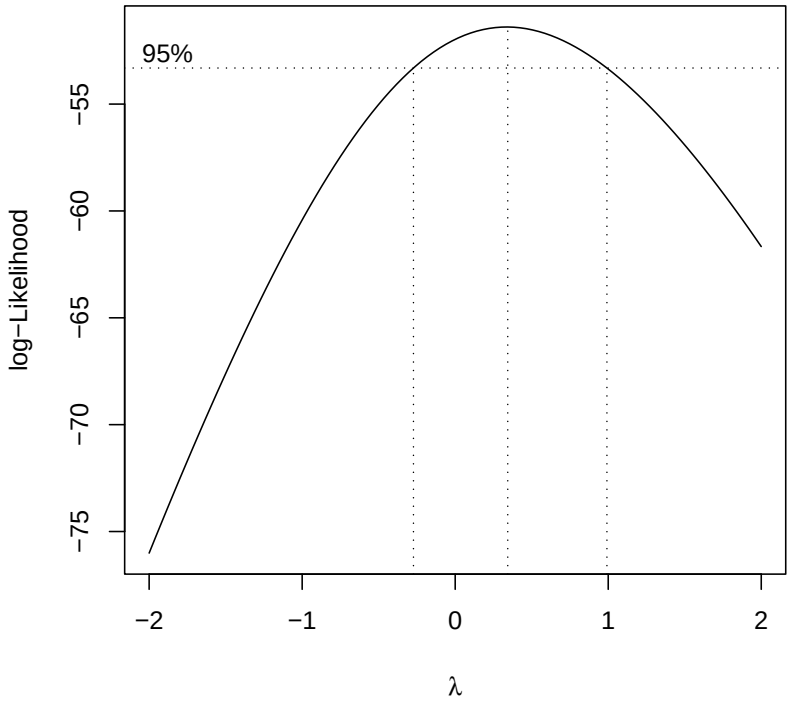

Figure 1: Output to `boxcox(modell.lm)`

2 Let Y_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, be random variables with

$$Y_{ijk} \;=\; \mu + \alpha_i + \beta_j + \epsilon_{ijk}\,, \tag{1}$$

where the ϵ_{ijk}'s are independent normally distributed random variables with mean zero and variance σ^2, $\sum_{i=1}^{I} \alpha_i = 0$ and $\sum_{j=1}^{J} \beta_j = 0$. By considering

$$S(\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J) \;=\; \sum_{i,j,k} (Y_{ijk} - \mu - \alpha_i - \beta_j)^2\,,$$

show that the least squares estimates of μ, α_i and β_j are respectively

$$\hat{\mu} \;=\; \bar{Y}_{+++}\,, \qquad \hat{\alpha}_i \;=\; \bar{Y}_{i++} - \bar{Y}_{+++} \qquad \text{and} \qquad \hat{\beta}_j \;=\; \bar{Y}_{+j+} - \bar{Y}_{+++}\,,$$

where $\bar{Y}_{i++} = \sum_{j,k} Y_{ijk}/(JK)$, $\bar{Y}_{+j+} = \sum_{i,k} Y_{ijk}/(IK)$ and $\bar{Y}_{+++} = \sum_{i,j,k} Y_{ijk}/(IJK)$.

Find the residual sum of squares RSS_1 for this model. Find the residual sum of squares RSS_0 for the null model $Y_{ijk} = \mu + \epsilon_{ijk}$. Show that the reduction in the residual sum of squares due to including the β_j's into the null model is the same as the reduction in the residual sum of squares due to including the β_j's into the model $Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$.

For a dataset with 3 levels for factor A, 2 levels for factor B, and with 2 replicates for each combination of factor levels, model (1) is fitted and the following analysis of variance table is obtained.

```
Analysis of Variance Table

Response: y
          Df   Sum Sq

A          *  12.7400
B          *   0.4033
Residuals  *   2.6867
```

where the degrees of freedom have been replaced by asterisks. What should the degrees of freedom be? Write down the values of the residual sums of squares for the following models (i) the null model, (ii) model (1), (iii) the model $Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$ and (iv) the model $Y_{ijk} = \mu + \beta_j + \epsilon_{ijk}$. How could you check whether there is an interaction between the factors A and B?

3 Let $Y_1, \ldots, Y_n$ be independent Poisson random variables, with $\mathbb{E}(Y_i) = \mu_i$. Assume that

$$\log(\mu_i) \,=\, \beta^T x_i \,, \quad i \,=\, 1 \,, \ldots \,, n \,,$$

where β is p-dimensional vector of parameters and x_i is a p-dimensional vector of known covariate values for the ith observation. Explain why this is a generalised linear model. Find the equations satisfied by the maximum likelihood estimator $\hat{\beta}$ of β based on observations $y_1, \ldots, y_n$. Find the deviance for this model. If the first component of x_i is 1 for all i, show that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i$, where $\hat{\mu}_i = \exp(\hat{\beta}^T x_i)$.

The number of different plant species was recorded for each of 90 plots, each with different biomass values. Thirty of the plots had low soil ph-level, thirty had medium soil ph-level and thirty had high soil ph-level. In the R commands below, `nspecies` contains the numbers of species for each of the plots, `biomass` contains the biomass values and `ph` is a factor with three levels (low, medium and high).

```
> model1.glm <- glm(nspecies ~ biomass * ph, poisson)
> model2.glm <- glm(nspecies ~ biomass + ph, poisson)
```

The deviances for `model1.glm` and `model2.glm` are 83.2 and 99.2 respectively. Write down in algebraic form the models that have been fitted, and illustrate with appropriate sketch graphs. What do you conclude about how the number of species depends on the biomass for the different soil ph-levels?

4 A statistician has data on the incidence of melanoma in women in two American cities, Minneapolis-Saint Paul in Minnesota and Forth Worth in Texas, for age groups, as shown in the (edited) R ouput below. In the output, `age` and `city` are factors giving respectively the age groups as shown and the city (0 is Minneapolis-Saint Paul and 1 is Fort Worth), `pop` and `cases` respectively contain the number in the population and the number of melanoma cases in the relevant age group and city. The statistician carries out two separate analyses, both using corner-point constraints.

(a) Comment on any obvious deficiencies in the data.

(b) State the model fitted in Analysis 1 and interpret the output in detail.

(c) What model is fitted in Analysis 2?

(d) Briefly compare and discuss the two analyses.

```
> propcases <- cases/pop

   cases  city    age     pop     propcases

1      1     0  15-24  172675  5.791226e-06
2     16     0  25-34  123065  1.300126e-04
3     30     0  35-44   96216  3.117985e-04
4     71     0  45-54   92051  7.713116e-04
5    102     0  55-64   72159  1.413545e-03
6    130     0  65-74   54722  2.375644e-03
7    133     0  75-84   32185  4.132360e-03
8     40     0    85+    8328  4.803074e-03
9      4     1  15-24  181343  2.205765e-05
10    38     1  25-34  146207  2.599055e-04
11   119     1  35-44  121374  9.804406e-04
12   221     1  45-54  111353  1.984679e-03
13   259     1  55-64   83004  3.120332e-03
14   310     1  65-74   55932  5.542444e-03
15    65     1    85+    7583  8.571805e-03

#  Analysis 1:

> melanoma1.glm <- glm(propcases ~ age + city, binomial, weights = pop)
> anova(melanoma1.glm, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit
Response: propcases
Terms added sequentially (first to last)

     Df  Deviance  Resid. Df  Resid. Dev  P(>|Chi|)
NULL                    14     2330.46
age   7   2098.19          7      232.28       0.00
city  1    227.12          6        5.15  2.526e-51
```

```
> summary(melanoma1.glm)

Call:
glm(formula = propcases ~ age + city, family = binomial, weights = pop)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2830  -0.3355   0.0000   0.3927   1.0820

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)

(Intercept) -11.69364     0.44923  -26.030   < 2e-16
age25-34      2.62915     0.46747    5.624  1.86e-08
age35-44      3.84627     0.45467    8.459   < 2e-16
age45-54      4.59538     0.45104   10.188   < 2e-16
age55-64      5.08901     0.45031   11.301   < 2e-16
age65-74      5.65031     0.44976   12.563   < 2e-16
age75-84      6.20887     0.45756   13.570   < 2e-16
age85+        6.18346     0.45783   13.506   < 2e-16
city1         0.85492     0.05969   14.322   < 2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2330.4637  on 14  degrees of freedom
Residual deviance:    5.1509  on  6  degrees of freedom

Number of Fisher Scoring iterations: 4

#  Analysis 2:

> melanoma2.glm <- glm(cases ~ offset(log(pop)) + age + city, poisson)
> anova(melanoma2.glm, test = "Chisq")

Analysis of Deviance Table

Model: poisson, link: log
Response: cases
Terms added sequentially (first to last)

     Df Deviance Resid.  Df  Resid. Dev  P(>|Chi|)
NULL                     14     2327.29
age   7  2095.56          7      231.73       0.00
city  1   226.52          6        5.21  3.423e-51
```

5 (a) Explain the following statistical terms used in survival analysis

(i) survival data,

(ii) right censoring,

(iii) left truncation.

(b) An academic researcher interested in Indian cuisine approaches two statisticians, Statistician A and Statistician B. The researcher has collected data on a sample of Indian restaurants, located in Cambridge and the surrounding villages, that were operational during all or part of the five-year period from the 1st January 2000 to 31st December 2005, and followed them up until the earlier of their date of closure and 31st December 2007.

The data collected comprise five variables:

obsage:	either the length of time the restaurant was in business, if closed down before the end of follow-up or the length of time the restaurant has been operating at the end of follow-up period, if still open
ageatentry:	the length of time the restaurant had been in operation at entry into the study.
status:	the closure status of the restaurant (**status = 1**, if closed down; **status = 0**, if still in business)
size:	the size of the restaurant (small, medium or large, coded 0, 1, and 2 respectively),
camb:	**camb = 1** if the restuarant is in Cambridge, and **camb = 0** if it is in a village.

All time variables are measured in years.

The academic is particularly interested in determining the proportion of Indian restaurants open for 5 years or more and open for 10 years or more, and asked the two statisticians if they could analyse the data collected. Both statisticians recognise that this is a survival data problem, and use the R statistical software environment to construct Kaplan-Meier curves. However, Statisticians A's and B's results for the 5-year and 10-year survival probabilities differ from one another, which confuses the researcher.

The researcher approaches you with the two statisticians' R codes and results, which are shown in the R output provided, and asks for your assistance.

(i) Examine the R output provided and determine, with explanation, which of the two statisticians has performed the more appropriate analysis for constructing the Kaplan-Meier curve. From your choice of the more appropriate analysis, what are the 5-year and 10-year survival probabilities? (The precision attached to point estimates is required.)

(ii) Also provided is the R output from a further analysis of the data performed by one of the statisticians. Comment in detail on the R commands, the analysis done and the results. (Derivations of the underlying techniques, e.g. for coxph( ), are not required.) What additional checks should be made before the results from this analysis are passed on to the academic researcher?

```
# Statistician A

> srvobj <- Surv(obsage, status)
> summary(survfit(srvobj~1))

Call: survfit(formula = srvobj ~ 1)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0.52	60	1	0.983	0.0165	0.951	1.000
0.63	59	1	0.967	0.0232	0.922	1.000
0.69	58	1	0.950	0.0281	0.896	1.000
0.84	57	1	0.933	0.0322	0.872	0.999
1.05	56	1	0.917	0.0357	0.849	0.989
1.32	55	1	0.900	0.0387	0.827	0.979
1.35	54	1	0.883	0.0414	0.806	0.968
1.48	53	1	0.867	0.0439	0.785	0.957
1.59	52	1	0.850	0.0461	0.764	0.945
1.65	51	1	0.833	0.0481	0.744	0.933
1.69	50	1	0.817	0.0500	0.724	0.921
1.86	49	1	0.800	0.0516	0.705	0.908
2.20	48	1	0.783	0.0532	0.686	0.895
2.21	47	1	0.767	0.0546	0.667	0.882
2.43	46	1	0.750	0.0559	0.648	0.868
2.68	45	1	0.733	0.0571	0.630	0.854
3.10	44	1	0.717	0.0582	0.611	0.840
3.31	43	1	0.700	0.0592	0.593	0.826
3.69	42	1	0.683	0.0601	0.575	0.812
4.40	41	1	0.667	0.0609	0.557	0.797
4.52	40	1	0.650	0.0616	0.540	0.783
5.24	38	1	0.633	0.0623	0.522	0.768
6.03	37	1	0.616	0.0629	0.504	0.752
6.47	36	1	0.599	0.0634	0.486	0.737
7.18	33	1	0.581	0.0641	0.468	0.721
7.30	32	1	0.562	0.0646	0.449	0.704
7.34	31	1	0.544	0.0650	0.431	0.688
7.42	30	1	0.526	0.0653	0.412	0.671
7.67	29	1	0.508	0.0655	0.394	0.654
7.76	26	1	0.488	0.0659	0.375	0.636
8.29	21	1	0.465	0.0667	0.351	0.616
9.19	16	1	0.436	0.0686	0.320	0.594
9.90	15	1	0.407	0.0699	0.291	0.570
10.58	13	1	0.376	0.0712	0.259	0.545
11.16	11	1	0.342	0.0724	0.225	0.518
12.42	6	1	0.285	0.0797	0.164	0.493

```
# Statistician B

> srvobj <- Surv(time = ageatentry, time2 = obsage,
                  event = status, type = "counting")

> summary(survfit(srvobj~1))

Call: survfit(formula = srvobj ~ 1)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0.52	25	1	0.960	0.0392	0.886	1.000
0.63	26	1	0.923	0.0523	0.826	1.000
0.69	26	1	0.888	0.0611	0.775	1.000
0.84	27	1	0.855	0.0671	0.733	0.997
1.05	30	1	0.826	0.0707	0.699	0.977
1.32	30	1	0.799	0.0735	0.667	0.957
1.35	29	1	0.771	0.0759	0.636	0.935
1.48	29	1	0.745	0.0778	0.607	0.914
1.59	31	1	0.721	0.0789	0.581	0.893
1.65	30	1	0.697	0.0799	0.556	0.872
1.69	29	1	0.672	0.0807	0.532	0.851
1.86	28	1	0.648	0.0813	0.507	0.829
2.20	28	1	0.625	0.0816	0.484	0.808
2.21	27	1	0.602	0.0818	0.461	0.786
2.43	27	1	0.580	0.0818	0.440	0.764
2.68	26	1	0.558	0.0816	0.419	0.743
3.10	26	1	0.536	0.0812	0.398	0.721
3.31	26	1	0.515	0.0807	0.379	0.701
3.69	26	1	0.496	0.0800	0.361	0.680
4.40	29	1	0.479	0.0790	0.346	0.661
4.52	29	1	0.462	0.0780	0.332	0.643
5.24	30	1	0.447	0.0769	0.319	0.626
6.03	34	1	0.434	0.0758	0.308	0.611
6.47	33	1	0.420	0.0746	0.297	0.595
7.18	30	1	0.406	0.0734	0.285	0.579
7.30	29	1	0.392	0.0722	0.274	0.563
7.34	28	1	0.378	0.0710	0.262	0.546
7.42	27	1	0.364	0.0697	0.250	0.530
7.67	27	1	0.351	0.0684	0.239	0.514
7.76	24	1	0.336	0.0671	0.227	0.497
8.29	21	1	0.320	0.0658	0.214	0.479
9.19	16	1	0.300	0.0647	0.197	0.458
9.90	15	1	0.280	0.0634	0.180	0.436
10.58	13	1	0.259	0.0621	0.162	0.414
11.16	11	1	0.235	0.0607	0.142	0.390
12.42	6	1	0.196	0.0620	0.105	0.364

```
# Further analysis performed on the "Indian restaurant" data

> size <- factor(size)
> size <- relevel(size,2)
> summary(coxph(srvobj ~ camb + size))

Call: coxph(formula = srvobj ~ camb + size)

 n = 60
        coef  exp(coef)  se(coef)       z     p

camb   -0.803     0.448     0.344  -2.335  0.02
size1  -0.322     0.724     0.482  -0.669  0.50
size3  -0.224     0.799     0.375  -0.598  0.55

       exp(coef) exp(-coef)  lower .95  upper .95

camb     0.448       2.23     0.228     0.879
size1    0.724       1.38     0.282     1.863
size3    0.799       1.25     0.383     1.666

Rsquare = 0.087   (max possible= 0.979)

Likelihood ratio test = 5.45  on 3 df,  p = 0.142
Wald test             = 5.76  on 3 df,  p = 0.124
Score (logrank) test  = 6.01  on 3 df,  p = 0.111
```

END OF PAPER