UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Tuesday, 1 June, 2010 9:00 am to 12:00 pm

PAPER 37

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

Suppose that $Y = (Y_1, \ldots, Y_n)^T$ satisfies $Y = X\beta + \varepsilon$, where X is a known $n \times p$ matrix with rank p $(< n)$, $\beta = (\beta_1, \ldots, \beta_p)^T$ is unknown, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ where $\varepsilon_1, \ldots, \varepsilon_n$ are independent normal random variables with mean zero and variance σ^2, and, where v^T denotes the transpose of v. Derive the least squares estimator $\hat{\beta}$ of β. Explain what is meant by the vector $\hat{Y}$ of fitted values and by the vector $\hat{\epsilon}$ of residuals. Find the distribution of $\hat{\epsilon}$. Show that $\hat{Y}$ is in the space spanned by the columns of X. Show that $X^T\hat{\epsilon} = 0$ and interpret this result.

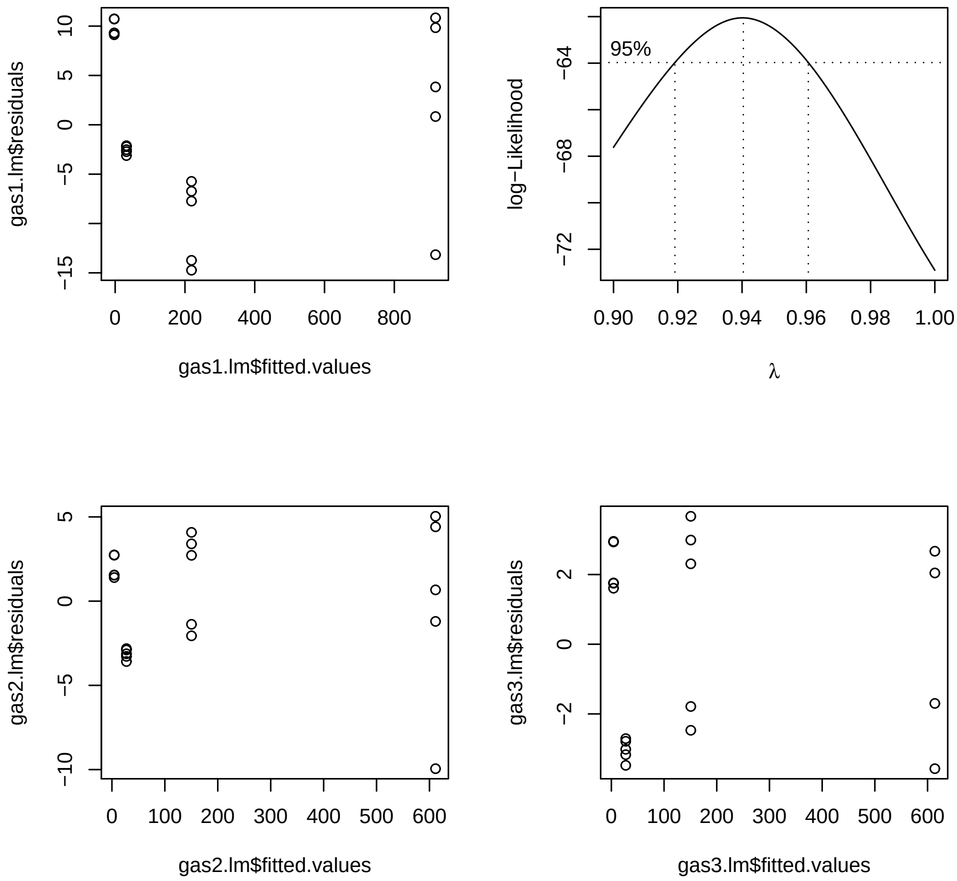
[*You may assume without proof that, for an* $m \times 1$ *random vector* W *and a* $k \times m$ *(constant) matrix* A, $\mathrm{cov}(AW) = A\mathrm{cov}(W)A^T$.]

Gas chromatography is a technique used to detect small amounts of a substance using a gas chromatograph. The edited R output below refers to a study in which five gas chromatograph readings were taken for each of four specimens containing different (known) amounts of the substance. The aim of the study is to calibrate the chromatograph by relating the actual amount of the substance to the chromatograph reading. In the R output `reading` contains the chromatograph readings and `amount` contains the amount of the substance in nanograms. The plots are also included below the output.

Write down the algebraic form of the model fitted in `gas1.lm`, together with any assumptions, and discuss whether or not this model seems to be satisfactory. Explain briefly what is shown in the `boxcox` plot and explain what you conclude from it. Write down the model fitted in `gas2.lm`. What features of the plot for this model might lead you to fit model `gas3.lm`? Using the `gas3.lm` model, explain how to obtain an estimate of the expected chromatograph reading when the amount of substance is 3.0 nanograms.

```
> gasdata
   amount  reading
1    0.25     6.55
2    0.25     7.98
3    0.25     6.54
4    0.25     6.37
5    0.25     7.96
6    1.00    29.70
7    1.00    30.00
8    1.00    30.10
9    1.00    29.50
10   1.00    29.10
11   5.00   211.00
12   5.00   204.00
13   5.00   212.00
14   5.00   213.00
15   5.00   205.00
16  20.00   929.00
17  20.00   905.00
18  20.00   922.00
19  20.00   928.00
20  20.00   919.00
> gas1.lm <- lm(reading~amount)
> plot(gas1.lm$fitted.values,gas1.lm$residuals)
> library(MASS)
```

```
> boxcox(gas1.lm,lambda=seq(0.9,1,0.02))
> gas2.lm <- lm(reading^0.94~amount)
> plot(gas2.lm$fitted.values,gas2.lm$residuals)
> gas2.lm$residuals
         1          2          3          4          5          6          7
 1.5509228  2.7444560  1.5425249  1.3996408  2.7278574 -3.1253835 -2.8953704
         8          9         10         11         12         13         14
-2.8187301 -3.2788029 -3.5858295  2.7195604 -2.0579993  3.4012854  4.0828176
        15         16         17         18         19         20
-1.3748961  5.0364763 -9.9464872  0.6688711  4.4126540 -1.2035679
> gas3.lm <- lm(reading[-17]^0.94 ~ amount[-17])
> summary(gas3.lm)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.62509    0.82284  -4.406 0.000387
amount[-17] 30.87295    0.08622 358.054  < 2e-16
Residual standard error: 2.832 on 17 degrees of freedom
Multiple R-Squared: 0.9999
> plot(gas3.lm$fitted.values,gas3.lm$residuals)
```



2

The table below shows car insurance premiums for various categories of policyholders with 0, 3, 6 or 9 points on their driving licenses. For each category of policyholder the top row gives the premiums for third party fire and theft only policies and the bottom row gives the premiums for comprehensive policies.

	Number of points			
	0	3	6	9
21 year old male	306	384	384	409
	500	555	555	605
21 year old female	266	304	279	287
	435	430	464	478
30 year old female	177	177	177	213
	320	325	325	268
40 year old male	154	162	162	189
	230	230	230	295

In the (edited) R output below, `Gender`, `Age`, `Policy` and `Points` are factors, and corner point constraints are used.

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

(b) Write down the algebraic form of the model fitted in `insurance1.lm`, defining your notation carefully and writing down the assumptions and constraints explicitly. You are given that the residual sum of squares for this model is 19512.

(c) You are given that the model `insurance2.lm` has residual sum of squares equal to 22323. What hypothesis is being tested by the test statistic whose value is `f`, and why does the test statistic take this form? What is the result of this hypothesis test? Write down your conclusion in words.

(d) Write down the algebraic form of the model fitted in `insurance3.lm`, again explicitly writing down the assumptions and constraints. Test whether this model is an improvement over `insurance2.lm`, and summarise in words how premiums depend on age, gender, policy type and the number of points. What is the estimated comprehensive policy premium for a 40 year old female policyholder with 6 points on her license?

```
> x
 [1] 306 384 384 409 500 555 555 605 266 304 279 287 435 430 464 478 177 177 177
[20] 213 320 325 325 368 154 162 162 189 230 230 230 295
> Gender
 [1] M M M M M M M M F F F F F F F F F F F F F F F F M M M M M M M M
Levels: F M
> Age
 [1] 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 30 30 30 30 30 30 30 30 40
[26] 40 40 40 40 40 40 40
Levels: 21 30 40
> Points
```

```
 [1] 0 3 6 9 0 3 6 9 0 3 6 9 0 3 6 9 0 3 6 9 0 3 6 9 0 3 6 9 0 3 6 9
Levels: 0 3 6 9
> Policy
 [1] 3rd  3rd  3rd  3rd  comp comp comp comp 3rd  3rd  3rd  3rd  comp comp comp
[16] comp 3rd  3rd  3rd  3rd  comp comp comp comp 3rd  3rd  3rd  3rd  comp comp
[31] comp comp
Levels: 3rd comp
> insurance1.lm  <- lm(x~Age+Gender+Policy+Points)

> Points2 <- factor(rep(c(1,1,1,2),times=8))
> Points2
 [1] 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 2
Levels: 1 2
> insurance2.lm <- lm(x~Age+Gender+Policy+Points2)
> f <- ((22323-19512)/2)/(19512/24)
> f
[1] 1.728782
> qf(0.95,2,24)
[1] 3.402826

> insurance3.lm <- lm(x~Age*Policy + Gender + Points2)
> anova(insurance3.lm)
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        2 275639  137820 329.850 < 2.2e-16
Policy     1 167476  167476 400.827 < 2.2e-16
Gender     1  35627   35627  85.267 2.276e-09
Points2    1  10438   10438  24.981 4.177e-05
Age:Policy  2  12295    6147  14.713 6.754e-05
Residuals  24  10028     418
> summary(insurance3.lm)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      269.760      9.094  29.665  < 2e-16
Age30            -94.187     13.520  -6.966 3.33e-07
Age40           -207.812     13.520 -15.370 6.38e-14
Policycomp       175.375     10.220  17.159 5.61e-15
GenderM           94.375     10.220   9.234 2.28e-09
Points22          41.708      8.345   4.998 4.18e-05
Age30:Policycomp  -26.875     17.702  -1.518    0.142
Age40:Policycomp  -95.875     17.702  -5.416 1.46e-05

Residual standard error: 20.44 on 24 degrees of freedom
Multiple R-Squared: 0.9804
F-statistic: 171.5 on 7 and 24 DF,  p-value: < 2.2e-16
```

3

Observations $Y_1, \ldots, Y_n$ are independent binary random variables with $\mathbb{P}(Y_i = 1) = p_i = 1 - \mathbb{P}(Y_i = 0)$, $i = 1, \ldots, n$. Assume that

$$\operatorname{logit}(p_i) \left(= \log\left(\frac{p_i}{1-p_i}\right)\right) = \beta^T x_i\,, \qquad i = 1, \ldots, n\,,$$

where β is a p-dimensional vector of unknown parameter values and x_i is a p-dimensional vector of known covariate values for the ith observation. Here β^T denotes the transpose of β.

(a) Show that $\mathbb{P}(Y_i = y_i)$ can be written in the form

$$\exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right),$$

and identify θ_i, $b(\theta_i)$ and ϕ.

(b) By considering the loglikelihood, derive an equation satisfied by the maximum likelihood estimator $\hat{\beta}$ of β. Let $p_i(\beta) = e^{\beta^T x_i}/(1 + e^{\beta^T x_i})$. Show that

$$\sum_{i=1}^{n} p_i(\hat{\beta}) \operatorname{logit}\left(p_i(\hat{\beta})\right) = \sum_{i=1}^{n} y_i \operatorname{logit}\left(p_i(\hat{\beta})\right).$$

(c) Show that deviance D can be expressed as

$$D = -2\sum_{i=1}^{n}\left(p_i(\hat{\beta})\operatorname{logit}\left(p_i(\hat{\beta})\right) + \log\left(1 - p_i(\hat{\beta})\right)\right).$$

Comment on the usefulness or otherwise of D as a measure of goodness fit in this case.

In a nature reserve in the United States there were 659 trees of a particular species before a storm, during which many of the trees were blown down. For each of the 659 trees, there is a record of the diameter T (in inches) and the severity of the storm at the tree's location, where the severity values are between 0 and 1, with higher values denoting higher severity. Suppose that `y` contains an indicator of whether or not the tree was blown down (`1` if the tree was blown down, `0` otherwise), `lT` contains $\log_2(T)$ for each tree, and `S` contains the severity value at each tree location. Write down the algebraic forms of the three models that would be fitted by the R directives

```
blow1.glm <- glm(y~1,binomial)
blow2.glm <- glm(y~lT,binomial)
blow3.glm <- glm(y~lT+S,binomial)
```

The deviances of the three models are 856.21, 655.24 and 563.90 respectively. Carry out a formal hypothesis test to determine whether `blow3.glm` is an improvement over `blow2.glm`. Using the (edited) R output below, give an expression for the estimated effect of doubling the diameter on the odds of a tree being blown down when the severity value is unchanged.

```
> summary(blow3.glm)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5621     0.7499  -12.75   <2e-16
lT            2.2164     0.2079   10.66   <2e-16
S             4.5086     0.5159    8.74   <2e-16
```

4

The (edited) R output below refers to a study into the effectiveness of some particular traffic control measures in reducing accident rates. In each of eight locations, there are data on the number of accidents over a number of years before and after the installation of the traffic control measures. In the R ouput below, `loc` contains the location identifiers (numbers between 1 and 8), `befaft` contains indicators of whether the observation was taken before or after installation (1 denotes before, 2 denotes afterwards), `years` contains the length of the observation period (in years), and `nacc` contains the number of accidents that occurred during that observation period. Corner point constraints are used.

(a) Explain what is calculated in line (*).

(b) Write down the algebraic form of the model fitted in `traffic1.glm`, defining your notation carefully and stating any assumptions. Using the output to `summary(traffic1.glm)`, show how to obtain an estimate of the ratio r of the accident rate after installation to the accident rate before installation. Explain how to obtain an approximate 95% confidence interval for r.

(c) Write down the algebraic form of the model in `traffic2.glm`. Why do you think this model is fitted? Comment on the fit of the model.

(d) Write a short paragraph giving relevant formal statistical analysis and your conclusions about the effect of the traffic measures on accident rates.

```
> loc
 [1] 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8
> befaft
 [1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
> years
 [1] 9 2 9 2 8 3 8 2 9 2 8 2 9 2 8 3
> nacc
 [1] 13  0  6  2 30  4 20  0 10  0 15  6  7  1 13  2
> Befaft <- factor(befaft)
> Loc <- factor(loc)
> r1 <- sum(nacc[befaft==1])/sum(years[befaft==1])
> r2 <- sum(nacc[befaft==2])/sum(years[befaft==2])
> r2/r1  # line (*)
[1] 0.497076
> traffic1.glm <- glm(nacc~offset(log(years))+Befaft,poisson)
> summary(traffic1.glm)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.51669    0.09366   5.517 3.45e-08
Befaft2     -0.69901    0.27466  -2.545   0.0109
    Null deviance: 58.589  on 15  degrees of freedom
Residual deviance: 50.863  on 14  degrees of freedom
> exp(-0.69901)
[1] 0.4970772
> traffic2.glm <- glm(nacc~offset(log(years))+Loc+Befaft,poisson)
> anova(traffic2.glm,test="Chisq")
       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      15     58.589
Loc     7   32.564         8     26.025 3.191e-05
```

```
Befaft  1    9.750         7     16.275     0.002
> summary(traffic2.glm)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2708     0.2785   0.972  0.33094
Loc2         -0.4855     0.4494  -1.080  0.27994
Loc3          1.0176     0.3264   3.117  0.00182
Loc4          0.5371     0.3563   1.507  0.13168
Loc5         -0.2624     0.4206  -0.624  0.53279
Loc6          0.5859     0.3529   1.660  0.09690
Loc7         -0.4855     0.4494  -1.080  0.27994
Loc8          0.1993     0.3792   0.526  0.59921
Befaft2      -0.7807     0.2754  -2.834  0.00459
    Null deviance: 58.589  on 15  degrees of freedom
Residual deviance: 16.275  on  7  degrees of freedom
```

5

A researcher has collected hospital data for swine influenza-related admissions during the middle period of the 2009 UK epidemic. Specifically, she has recorded the dates of admission, swine influenza-related death and discharge, and the time still in hospital since admission if a patient has yet to be discharged or to die from swine influenza-related causes at the time of data collection. She approaches you with the data and is particularly interested in the case fatality ratio θ associated with hospitalisation (i.e. the proportion of swine influenza-related hospital cases who eventually die from the disease) and the conditional distribution corresponding to the time of death given that a case will eventually die ($I = 1$) from swine influenza-related causes (with distribution function $F(t|I = 1)$ and density $f(t|I = 1)$). The conditional distribution corresponding to the time to recovery (i.e. discharge) given that a case will eventually recover ($I = 2$) from the illness (with distribution function $F(t|I = 2)$ and density $f(t|I = 2)$) may also be of interest. You recognise that this is a survival analysis problem and offer to help her analyse the data.

By appropriately defining all notation used:

(a) Identify which type(s) of patients correspond to right-censored observations.

(b) Write down the likelihood contributions for a case (i.e. a swine influenza-related admitted patient) who

  (i) dies in hospital at time t after admission;

  (ii) recovers and is discharged at time t after admission;

  (iii) remains in hospital at time t after admission.

(c) Derive an E-M algorithm, giving full details for the E-step, that can be used to estimate the parameters of interest to the researcher given that the conditional densities, $f(t|I = 1)$ and $f(t|I = 2)$, associated with time to swine influenza-related death and time to recovery given eventual death from swine influenza-related causes and eventual recovery respectively, are log-normal densities with parameters (μ_1, σ_1) and (μ_2, σ_2).

[*Hint: if X has a log-normal distribution with parameter (μ, σ), then $Y = \log(X)$ has a normal distribution with mean μ and variance σ^2. Also, if Y has a $N(\mu, \sigma^2)$ distribution, then, writing $z = (y - \mu)/\sigma$, we have $E(Y|Y > y) = \mu + \sigma\psi(z)$,*

$$E\left(\left(\frac{Y-a}{b}\right)^2 \Big| Y > y\right) = \frac{1}{b^2}\left\{\sigma^2[1-\omega(z)] + [(\mu - a) + \sigma\psi(z)]^2\right\}$$

for constants a and b ($\neq 0$), and

$$var(Y|Y>0) = \sigma^2[1-\omega(z)] \quad where \quad \psi(z) = \frac{\phi(z)}{1-\Phi(z)} \quad and \quad \omega(z) = \psi(z)[\psi(z) - z],$$

and where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function respectively for a standard normal distribution.]

END OF PAPER