Student Number

Semester 2 Assessment, 2017

School of Mathematics and Statistics

MAST30027 Modern Applied Statistics

Writing time: 3 hours

Reading time: 15 minutes

This is NOT an open book exam

This paper consists of 8 pages (including this page)

Authorised Materials

- Mobile phones, smart watches and internet or communication devices are forbidden.
- A single two-sided hand-written A4 sheet of notes.
- The only permitted scientific calculator is the Casio FX82.

Instructions to Students

- You must NOT remove this question paper at the conclusion of the examination.
- You should attempt all questions. Marks for individual questions are shown.
- The total number of marks available is 120.

Instructions to Invigilators

- Students must NOT remove this question paper at the conclusion of the examination.

Blank page (ignored in page numbering)

Question 1 (10 marks) Let $X_1, \cdots, X_n$ be independent samples from a Pareto distribution $\text{Par}(1, \kappa)$ with pdf $f(x|\kappa) = \kappa(1+x)^{-\kappa-1}, x > 0$.

(a) What is the log-likelihood for this example?

(b) What is the Fisher information for this example?

(c) Find the MLE of κ and its asymptotic distribution.

Question 2 (8 marks) The gamma distribution with shape parameter $\nu > 0$ and rate parameter $\lambda > 0$ has the probability density function

$$f(x; \nu, \lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}, x \geq 0.$$

The mean is $\frac{\nu}{\lambda}$ and the variance is $\frac{\nu}{\lambda^2}$.

(a) Show that the gamma distribution is an exponential family.

(b) Obtain the canonical link. Show your work.

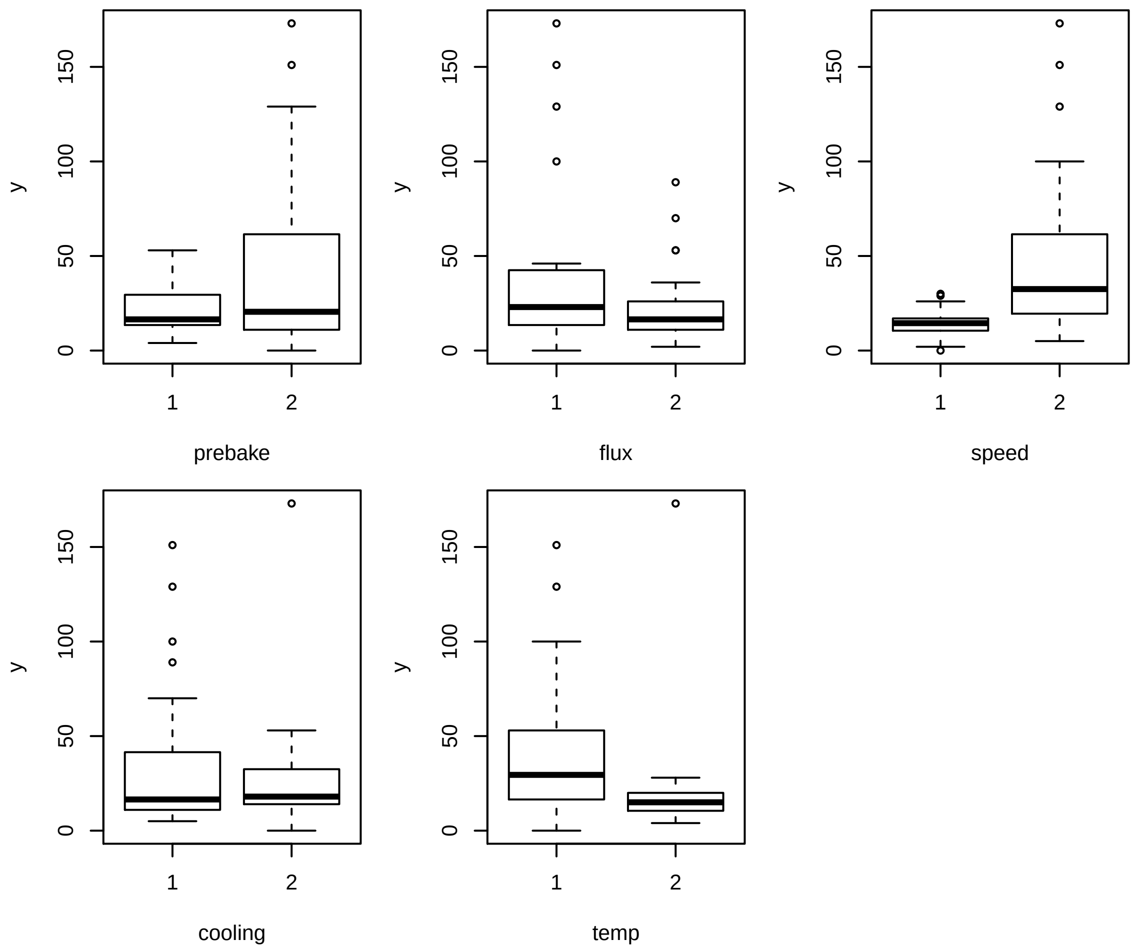
(c) Obtain the variance function. Show your work.

Question 3 (22 marks) The `wavesolder` data set has 48 observations of `y`, the number of defects, and five predictor variables, `prebake`, `flux`, `speed`, `cooling`, and `temp`. The data is taken from Condra, Lloyd, *Reliability Improvement with Design of Experiment*, CRC Press, 2001.

Examine the R code and output below, and then answer the questions that follow.

Firstly we need to combine the three replicates into a single data set, and then have a look at the data.

```
> rm(list=ls())
> library(faraway)
> data(wavesolder)
> y <- c(wavesolder$y1, wavesolder$y2, wavesolder$y3)
> wavesolder <- rbind(wavesolder, wavesolder, wavesolder)
> wavesolder <- wavesolder[-(1:3)]
> wavesolder$y <- y
> par(mfrow=c(2,3), mar=c(4,4,1,1))
> plot(y ~ prebake, wavesolder)
> plot(y ~ flux, wavesolder)
> plot(y ~ speed, wavesolder)
> plot(y ~ cooling, wavesolder)
> plot(y ~ temp, wavesolder)
```



```
> modelA <- glm(y ~ prebake + flux + speed + temp,
+              family=poisson, data=wavesolder)
> summary(modelA)
Call:
glm(formula = y ~ prebake + flux + speed + temp, family = poisson,
    data = wavesolder)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.0503  -1.9044  -0.5489   1.8995  12.5918

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.80541    0.06948   40.38   <2e-16 ***
prebake2     0.67287    0.05374   12.52   <2e-16 ***
flux2       -0.52878    0.05262  -10.05   <2e-16 ***
speed2       1.23048    0.06076   20.25   <2e-16 ***
temp2       -0.69315    0.05392  -12.86   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1450.52  on 47  degrees of freedom
Residual deviance:  513.75  on 43  degrees of freedom
AIC: 754.49

Number of Fisher Scoring iterations: 5

> modelB <- glm(y ~ prebake + flux + speed + cooling + temp,
+              family=poisson, data=wavesolder)
> summary(modelB)
Call:
glm(formula = y ~ prebake + flux + speed + cooling + temp, family = poisson,
    data = wavesolder)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.7230  -2.0135  -0.2761   1.5991  13.2687

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.88947    0.07576  38.142  < 2e-16 ***
prebake2     0.64801    0.05450  11.891  < 2e-16 ***
flux2       -0.52878    0.05262 -10.049  < 2e-16 ***
speed2       1.21614    0.06098  19.943  < 2e-16 ***
cooling2    -0.14222    0.05279  -2.694  0.00706 **
temp2       -0.66902    0.05463 -12.247  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1450.52  on 47  degrees of freedom
Residual deviance:  506.48  on 42  degrees of freedom
AIC: 749.22

Number of Fisher Scoring iterations: 5

> anova(modelA, modelB, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ prebake + flux + speed + temp
Model 2: y ~ prebake + flux + speed + cooling + temp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        43     513.75
2        42     506.48  1   7.2729    0.007 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


> (phi <- sum(residuals(modelB, type="pearson")^2/modelB$df.residual))
[1] 13.93209

> modelC <- glm(y ~ prebake + flux + speed + temp, family=quasipoisson, data=wavesolder)
> modelD <- glm(y ~ prebake + flux + speed + cooling + temp, family=quasipoisson,
+ data=wavesolder)
> anova(modelC, modelD, test="F")
Analysis of Deviance Table

Model 1: y ~ prebake + flux + speed + temp
Model 2: y ~ prebake + flux + speed + cooling + temp
  Resid. Df Resid. Dev Df Deviance     F Pr(>F)
1        43     513.75
2        42     506.48  1   7.2729 0.522  0.474
```

(a) For `modelA`, assuming Poisson responses, what is the log-likelihood of the fitted model, and the log-likelihood of the full (saturated) model?

(b) Assuming Poisson responses, which is better, `modelA` or `modelB`? Give two (quantitative) reasons for your answer.

(c) Give an estimate for the expected number of defects, for `prebake = 1, flux = 2, speed = 2, cooling = 1, temp = 1`, under `modelB`.

(d) Give a (quantitative) reason why `modelB` may suffer from overdispersion.

(e) Briefly the difference between a Poisson and quasi-Poisson model.

(f) Give the std. error for `cooling2` in the case where we allow for overdispersion.

(g) Allowing for overdispersion, do you prefer `modelC` or `modelD`, and why?

(h) What formula has been used to calculate the F statistic in the second analysis of deviance? What are the degrees of freedom for the F statistic?

Question 4 (14 marks) The following three-way table refers to results of a case-control study about effects of cigarette smoking and coffee drinking on myocardial infarction (MI) or heart attack for a sample of men under 55 years of age.

	Cigarettes per Day							
Cups Coffee	0		1-24		25-34		≥ 35	
per Day	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
0	66	123	30	52	15	12	36	13
1-2	141	179	59	45	53	22	69	25
3-4	113	106	63	65	55	16	119	30
≥ 5	129	80	102	58	118	44	373	85

Eight log-linear models with Poisson error have been fitted, with the residual deviances given in the following table.

Model	Residual deviance
`coffee + cigar + MI`	607.25
`coffee + cigar*MI`	394.43
`cigar + coffee*MI`	484.70
`MI + coffee*cigar`	271.40
`coffee*cigar + coffee*MI`	148.81
`coffee*cigar + cigar*MI`	58.55
`coffee*MI + cigar*MI`	271.88
`coffee*cigar + coffee*MI + cigar*MI`	11.17

You will find the following chi-squared percentage points useful for problems (c) and (d).

```
> qchisq(0.95, df=5:10)
[1] 11.07050 12.59159 14.06714 15.50731 16.91898 18.30704
> qchisq(0.95, df=11:15)
[1] 19.67514 21.02607 22.36203 23.68479 24.99579
> qchisq(0.95, df=16:20)
[1] 26.29623 27.58711 28.86930 30.14353 31.41043
```

(a) What are the residual degrees of freedom (d.f.) for each of the three models: `coffee + cigar + MI`, `cigar + coffee*MI`, and `coffee*cigar + cigar*MI`?

(b) Give an interpretation for each of the following models.

(i) `coffee + cigar + MI`

(ii) `MI + coffee*cigar`

(iii) `coffee*cigar + coffee*MI`

(c) Test the hypothesis that there is no association between `coffee` and `MI` when `cigar` level is given (at the 95% level).

(d) Test the hypothesis that the association between `MI` and `cigar` is the same for all `coffee` levels. That is, test that there is no three-way interaction (at the 95% level).

Question 5 (14 marks)

(a) Here is some R code for simulating a discrete random variable Y. What is the probability mass function (pmf) of Y, i.e., $P(Y = y)$ for $y \geq 2$?

```
Y.sim <- function() {
  U <- runif(1)
  Y <- 2
  while (U > 1 - 1/Y) {
    Y <- Y + 1
  }
  return(Y)
}
```

(b) Let a random number X be generated by the following algorithm:

1° Generate U from $\text{Unif}(0, 1)$ and V from $\text{Unif}(0, 1)$ independently.

2° If $U + V < 1$, then $X = 1 - U$; otherwise, go to 1°.

What is the probability density function of X, i.e., $f(x)$ for $0 \leq x \leq 1$?

Question 6 (18 marks) Consider a random sample X from a Bernoulli distribution with pdf $f(x|\theta) = \theta^x(1-\theta)^{1-x};\ \ x = 0, 1$. Let the prior distribution for θ be $\text{Uniform}(0, 1)$, i.e., $p(\theta) = 1$ for $0 < \theta < 1$. We use the squared error loss function.

(a) Find the posterior distribution of θ.

(b) Find the Bayes estimator of θ.

(c) Find the risk of the Bayes estimator of θ.

(d) Find the Bayes risk of the Bayes estimator of θ.

Question 7 (20 marks) We assume that $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ are independently normally distributed as follows.

$$\begin{aligned} x_i &\sim N(\mu_1, \sigma^2), \quad i = 1, \ldots, n_1 \\ y_i &\sim N(\mu_2, \sigma^2), \quad i = 1, \ldots, n_2 \end{aligned}$$

We impose the following prior distributions on μ_1, μ_2 and $\tau = 1/\sigma^2$.

$$\begin{aligned} p(\mu_1) &\propto 1 \\ p(\mu_2) &\propto 1 \\ p(\tau) &\propto 1/\tau \end{aligned}$$

(a) Among μ_1, μ_2 and τ, which parameter(s) have an improper prior?

(b) Write down the kernels of

$$\begin{aligned} &\mu_1 | \mathbf{x}, \mathbf{y}, \mu_2, \tau \\ &\mu_2 | \mathbf{x}, \mathbf{y}, \mu_1, \tau \\ &\tau | \mathbf{x}, \mathbf{y}, \mu_1, \mu_2 \end{aligned}$$

Hence give the (conditional) distributions of these variables, including their parameters.

(c) Briefly describe a Gibbs sampling scheme for sampling (μ_1, μ_2, τ).

(d) How would you check for convergence of the Gibbs sampler? Provide both informal (graphical/visual checks) and formal methods. Also, briefly provide details of methods.

Question 8 (14 marks) Briefly describe an algorithm to simulate samples from the posterior predictive distribution:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta) p(\theta|y) d\theta.$$

How would you estimate the mean of the posterior predictive distribution using the simulated samples?

End of Exam—Total Available Marks = 120