THE UNIVERSITY OF MELBOURNE

Student Number

Semester 2 Assessment, 2018

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 10 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.
- There are 8 questions with marks as shown. The total number of marks available is 100.

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) The Poisson distribution has the probability density function (pdf)

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ for } x = 0, 1, \ldots$$

(a) Show that the Poisson distribution is an exponential family, identifying the parameters θ and ϕ as well as the functions $b(\theta)$ and $a(\phi)$.

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

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

Question 2 (8 marks) Let $X_1, \cdots, X_n$ be independent random variables from a Poisson distribution with the pdf given in Question 1.

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

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

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

Question 3 (18 marks) The Fiji Fertility Survey gives data on the number of children ever born to married women of the Indian race classified by duration since their first marriage (grouped in six categories), type of place of residence (Suva, other urban and rural), and educational level (classified in four categories: none, lower primary, upper primary, and secondary or higher). The dataset has 70 rows representing 70 groups of families. Each row has 5 entries giving:

- duration: marriage duration of mothers in the group (years),
- residence: residence of families in the group (Suva, urban, rural),
- education: education of mothers in the group (none, lower primary, upper primary, secondary+),
- nChildren: number of children ever born in the group (e.g. 4), and
- nMother: number of mothers in the group (e.g. 8).

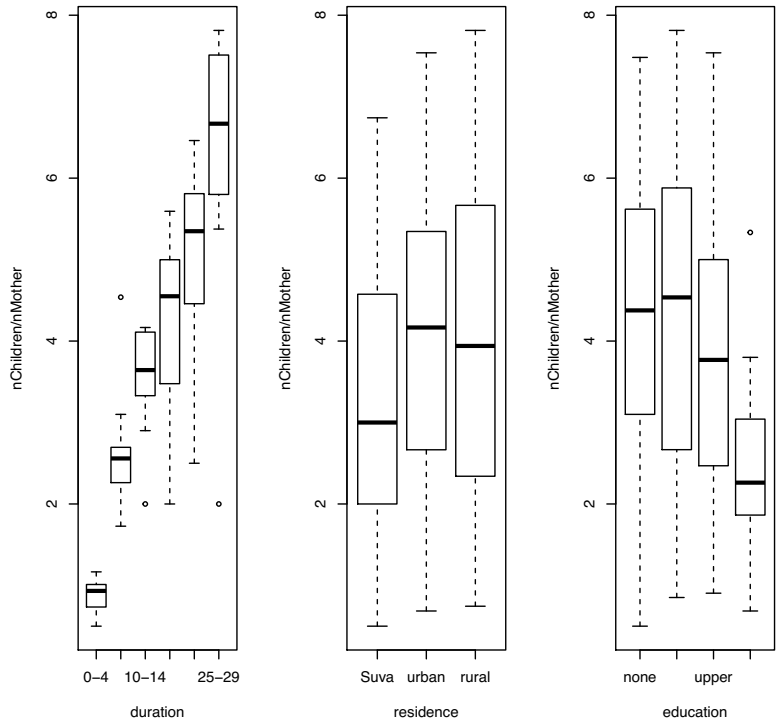
Examine the R code and output below, and then answer the questions that follow. First, we can summarise data as a table as follows.

```
> dat <- read.table("examData.dat", header=TRUE)
> dat$duration <- factor(dat$duration, levels=c("0-4","5-9","10-14","15-19","20-24",
"25-29"), ordered=TRUE)
> dat$residence <- factor(dat$residence, levels=c("Suva", "urban", "rural"))
> dat$education <- factor(dat$education, levels=c("none", "lower", "upper", "sec+"))
> ftable(xtabs(cbind(nChildren,nMother) ~ duration + residence + education, dat))

                            nChildren nMother
duration residence education
0-4      Suva      none             4       8
                   lower           24      21
                   upper           38      42
                   sec+            37      51
          urban    none            14      12
                   lower           23      27
                   upper           41      39
                   sec+            35      51
          rural    none            60      62
                   lower           98     102
                   upper          104     107
                   sec+            35      47
5-9       Suva     none            31      10
                   lower           80      30
                   upper           49      24
                   sec+            38      22
          urban    none            59      13
                   lower           98      37
                   upper          118      44
                   sec+            48      21
          rural    none           171      70
                   lower          317     117
                   upper          200      81
                   sec+            47      21
10-14     Suva     none            49      12
```

		lower	99	27
		upper	58	20
		sec+	24	12
	urban	none	75	18
		lower	143	43
		upper	105	29
		sec+	50	15
	rural	none	364	88
		lower	546	132
		upper	197	50
		sec+	30	9
15-19	Suva	none	59	14
		lower	153	31
		upper	41	13
		sec+	11	4
	urban	none	108	23
		lower	225	42
		upper	92	20
		sec+	19	5
	rural	none	577	114
		lower	481	86
		upper	135	30
		sec+	2	1
20-24	Suva	none	118	21
		lower	91	18
		upper	47	12
		sec+	13	5
	urban	none	118	22
		lower	147	25
		upper	65	13
		sec+	16	3
	rural	none	756	117
		lower	431	68
		upper	132	23
		sec+	5	2
25-29	Suva	none	310	47
		lower	182	27
		upper	43	8
		sec+	2	1
	urban	none	300	46
		lower	338	45
		upper	98	13
		sec+	0	0
	rural	none	1459	195
		lower	461	59
		upper	58	10
		sec+	0	0

```
> par(mfrow=c(1,3))
> plot(nChildren/nMother ~ duration, dat)
> plot(nChildren/nMother ~ residence, dat)
> plot(nChildren/nMother ~ education, dat)
```




```
> modelA <- glm(nChildren ~ offset(log(nMother)) + duration + residence
+ education, dat, family=poisson)
> summary(modelA)
Call:
glm(formula = nChildren ~ offset(log(nMother)) + duration + residence +
    education, family = poisson, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2960  -0.6641   0.0725   0.6336   3.6782

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.17314    0.03054  38.415  < 2e-16 ***
duration.L      1.49288    0.03387  44.082  < 2e-16 ***
duration.Q     -0.52726    0.03026 -17.424  < 2e-16 ***
duration.C      0.25258    0.02776   9.098  < 2e-16 ***
duration^4     -0.07613    0.02570  -2.962 0.003059 **
duration^5      0.03025    0.02402   1.259 0.207880
residenceurban  0.11242    0.03250   3.459 0.000541 ***
residencerural  0.15166    0.02833   5.353 8.63e-08 ***
educationlower  0.02297    0.02266   1.014 0.310597
educationupper -0.10127    0.03099  -3.268 0.001082 **
educationsec+  -0.31015    0.05521  -5.618 1.94e-08 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3731.852  on 69  degrees of freedom
Residual deviance:   70.665  on 59  degrees of freedom
AIC: 522.14

Number of Fisher Scoring iterations: 4

> modelB <- glm(nChildren ~ offset(log(nMother)) + duration + residence,
dat, family=poisson)
> summary(modelB)
Call:
glm(formula = nChildren ~ offset(log(nMother)) + duration + residence,
    family = poisson, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7517  -1.1748  -0.3193   0.8237   4.1255

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.11743    0.02562  43.611  < 2e-16 ***
duration.L      1.56800    0.03148  49.813  < 2e-16 ***
duration.Q     -0.55316    0.02997 -18.460  < 2e-16 ***
duration.C      0.26194    0.02772   9.448  < 2e-16 ***
duration^4     -0.07149    0.02568  -2.783 0.005383 **
duration^5      0.03528    0.02400   1.470 0.141478
residenceurban  0.11895    0.03246   3.664 0.000248 ***
residencerural  0.18192    0.02785   6.532  6.5e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3731.85  on 69  degrees of freedom
Residual deviance:  120.68  on 62  degrees of freedom
AIC: 566.16

Number of Fisher Scoring iterations: 4

> anova(modelB, modelA, test="Chisq")

Analysis of Deviance Table

Model 1: nChildren ~ offset(log(nMother)) + duration + residence
Model 2: nChildren ~ offset(log(nMother)) + duration + residence + education
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        62    120.680
```

```
2        59     70.665  3   50.015 7.93e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> (phi<-sum(residuals(modelA, type="pearson")^2/modelA$df.residual))
[1] 1.212432

> modelC <- glm(nChildren ~ offset(log(nMother)) + duration + residence
+ education, dat, family=quasipoisson)
> modelD <- glm(nChildren ~ offset(log(nMother)) + duration + residence,
dat, family=quasipoisson)
> anova(modelD, modelC, test="F")

Analysis of Deviance Table

Model 1: nChildren ~ offset(log(nMother)) + duration + residence
Model 2: nChildren ~ offset(log(nMother)) + duration + residence + education
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        62    120.680
2        59     70.665  3   50.015 13.751 6.573e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(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 the std. error for `educationlower` in the case where we allow for overdispersion.

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

(e) 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 (15 marks) The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The dataset can be found in the the dataset `pima`. The dataset contains the following variables:

- test: test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)
- pregnant: Number of times pregnant
- glucose: Plasma glucose concentration at 2 hours in an oral glucose tolerance test
- diastolic: Diastolic blood pressure (mm Hg)
- triceps: Triceps skin fold thickness (mm)
- insulin: 2-Hour serum insulin (mu U/ml)
- bmi: Body mass index (weight in kg/(height in metres squared))
- diabetes: Diabetes pedigree function
- age: Age (years)

Examine the R code and output below, and then answer the questions that follow.
There are missing observations for many variables, which have all been recorded as zeros. These are removed first.

```
> library("faraway")
> data(pima)
> missing <- with(pima, missing <- glucose==0 | diastolic==0 | triceps==0 | bmi == 0)
> pima <- pima[!missing,]
> model <- glm(cbind(test, 1-test)~., family=binomial, data=pima)
> model2 <- step(model, scope=~., trace=0)
> summary(model2)

Call:
glm(formula = cbind(test, 1 - test) ~ pregnant + glucose + bmi +
    diabetes + age, family = binomial, data = pima)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9894  -0.6530  -0.3700   0.6442   2.5417

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.879992   0.918614 -10.755  < 2e-16 ***
pregnant     0.123887   0.043514   2.847 0.004412 **
glucose      0.035026   0.004201   8.338  < 2e-16 ***
bmi          0.085123   0.018110   4.700  2.6e-06 ***
diabetes     1.321554   0.362538   3.645 0.000267 ***
age          0.023844   0.013311   1.791 0.073244 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

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

```
    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 467.08  on 526  degrees of freedom
AIC: 479.08

Number of Fisher Scoring iterations: 5
```

(a) Under the model2, give an estimate of the probability of the patient showing positive signs of diabetes when `pregnant` $= 1$, `glucose` $= 99$, `bmi` $= 27$, `diabetes` $= 0.25$, `age` $= 25$.

(b) Odds are sometimes a better scale than probability to represent chance. The odds o and probability p are related by

$$o = \frac{p}{1-p} \quad p = \frac{o}{1+o}$$

In a binomial regression model with a logit link we have

$$\text{logit}(p_j) = \log\left(\frac{p_j}{1-p_j}\right) = \eta_j = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_q x_{q,j}.$$

That is $\log o_j = \eta_j$, where o_j are the odds for the j-th observation.

By what proportion do the odds of testing positive for diabetes change for a woman with a BMI at 27.87 compared with a woman with a BMI at 36.90, assuming that all other factors are held constant? Give a 95% confidence interval for this difference.

Question 5 (18 marks) Consider a random sample $X_1, \cdots, X_n$ satisfying $X_i \stackrel{d}{=} \text{pois}(\theta)$, i.e., $f(x_i|\theta) = \frac{\theta^{x_i} e^{-\theta}}{x_i!}$. To assess an estimator $\hat{\theta} = t(X_1, \cdots, X_n)$ of θ we use the loss function $L(\hat{\theta}; \theta) = \dfrac{(\hat{\theta} - \theta)^2}{\theta}$. We assume a gamma prior pdf $\theta \stackrel{d}{=} \text{gamma}(\beta, \kappa)$ with known β and κ, i.e. $p(\theta) = \dfrac{1}{\beta^\kappa \Gamma(\kappa)} \theta^{\kappa-1} e^{-\theta/\beta};\ \theta > 0$.

(a) Show that the Bayes estimator under the given loss function is $\hat{\theta} = (E[\theta^{-1}|\mathbf{x}])^{-1}$.

(b) What is the posterior distribution of θ. Show your work.

(c) Find a closed form for $\hat{\theta}$ in (a) by using the result of (b). Show your work. [Hint: $\Gamma(n) = (n-1)!$].

Question 6 (7 marks) You only have access to uniform random number generator. For a random X, $f(x) \propto \log(1+x)$ if $x \in [0,2]$, and 0 else. Provide a rejection sampling algorithm (or pseudocode) to sample X. Your algorithm should be the most efficient algorithm (i.e., one that minimises the probability of rejection).

Question 7 (6 marks) Consider a Metropolis-Hastings (MH) algorithm where you propose θ_n from a Normal(θ_o, 1) distribution (where θ_o is the current sample) and accept with probability $\min(1, \frac{\exp(\frac{\theta_o^2}{2})}{\exp(\frac{\theta_n^2}{2})})$. What is the stationary distribution? Briefly describe the MH algorithm to simulate samples from that stationary distribution.

Question 8 (18 marks) You want to sample from $f(\theta, \mu|x) \propto \exp(-\frac{(\theta-\mu-1)^2}{(\mu+1)x^2})$, where θ is real-valued and $\mu \in \{0,1\}$ and x is known.

(a) Give the conditional distributions of $\theta|\mu, x$ and $\mu|\theta, x$, including their parameters.

(b) Briefly describe a Gibbs sampling algorithm for sampling (θ, μ), and how you would use it to estimate the mean of $\frac{\theta}{\mu}$.

End of Exam—Total Available Marks = 100