THE UNIVERSITY OF MELBOURNE

Student Number

Semester 2 Assessment, 2016

Department 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 9 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.
- Scientific calculators. Graphical calculators are *not* allowed.

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 80.

Instructions to Invigilators

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

Blank page (ignored in page numbering)

Question 1 (8 marks) Suppose that $Y \sim \text{negbin}(r, p)$ for some known r and unknown p. That is (taking r to be integer valued)

$$\mathbb{P}(Y = y) = \binom{y + r - 1}{r - 1}(1 - p)^r p^y, \quad \mathbb{E}Y = \frac{pr}{1 - p}.$$

If p is the probability of success, then Y is the number of successful trials until the r-th failure.

(a) What are the log-likelihood, observed information, and Fisher information for this example?

(b) Suppose that $r = 2$ and $y = 4$. Perform two steps of the Fisher scoring algorithm for maximising the log-likelihood, starting at $p(0) = 0.5$. Show your working.

(c) What is the asymptotic approximation to the distribution of $\hat{p}$, the MLE estimator of p?

Question 2 (6 marks) The Weibull distribution, with shape $k > 0$ and rate $\lambda > 0$, has probability density function

$$f(x) = kx^{k-1}\lambda^k e^{-(\lambda x)^k} \text{ for } x \geq 0.$$

We will assume that k is known and fixed. The mean is $\Gamma(1 + 1/k)/\lambda$ and the variance is $(\Gamma(1 + 2/k) - \Gamma(1 + 1/k)^2)/\lambda^2$.

(a) Suppose that $X \sim \text{Weibull}(k, \lambda)$, then find a transformation $Y = g(X)$ such that Y is from an exponential family.

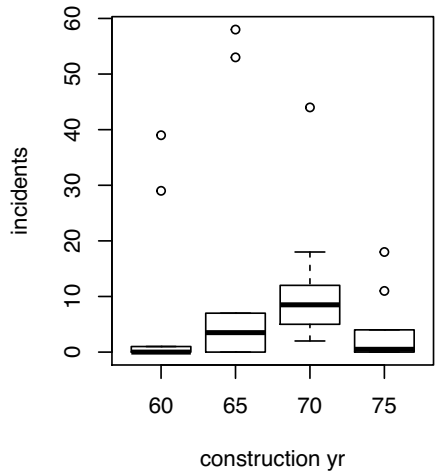
(b) Could you fit a model with Weibull responses using quasi-likelihood? If so, how?

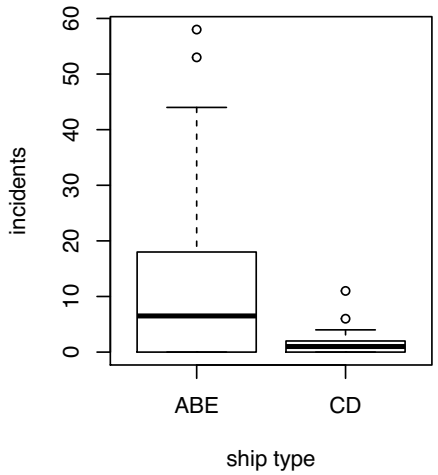
Your answer should be at most half a page.

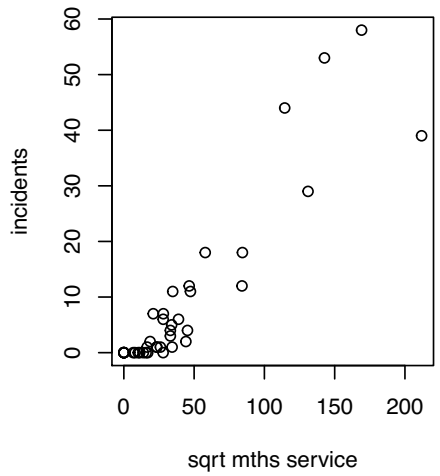
Question 3 (20 marks) The `ships` dataset from the `MASS` package gives the number of damage incidents and aggregate months of service for different types of ships, broken down by year of construction.

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

```
> library(MASS)
> data(ships)
> ships$year <- factor(ships$year)
> ships$type <- sapply(ships$type, function(x) if (x %in% c("A", "B", "E")) "ABE" else "CD")
> ships$type <- factor(ships$type)
> ships$rootserv <- sqrt(ships$service)
> par(mfrow=c(2,2))
> par(mar=c(4,4,1,2))
> plot(ships$year, ships$incidents, xlab="construction yr", ylab="incidents")
> plot(ships$type, ships$incidents, xlab="ship type", ylab="incidents")
> plot(ships$rootserv, ships$incidents, xlab="sqrt mths service", ylab="incidents")
```





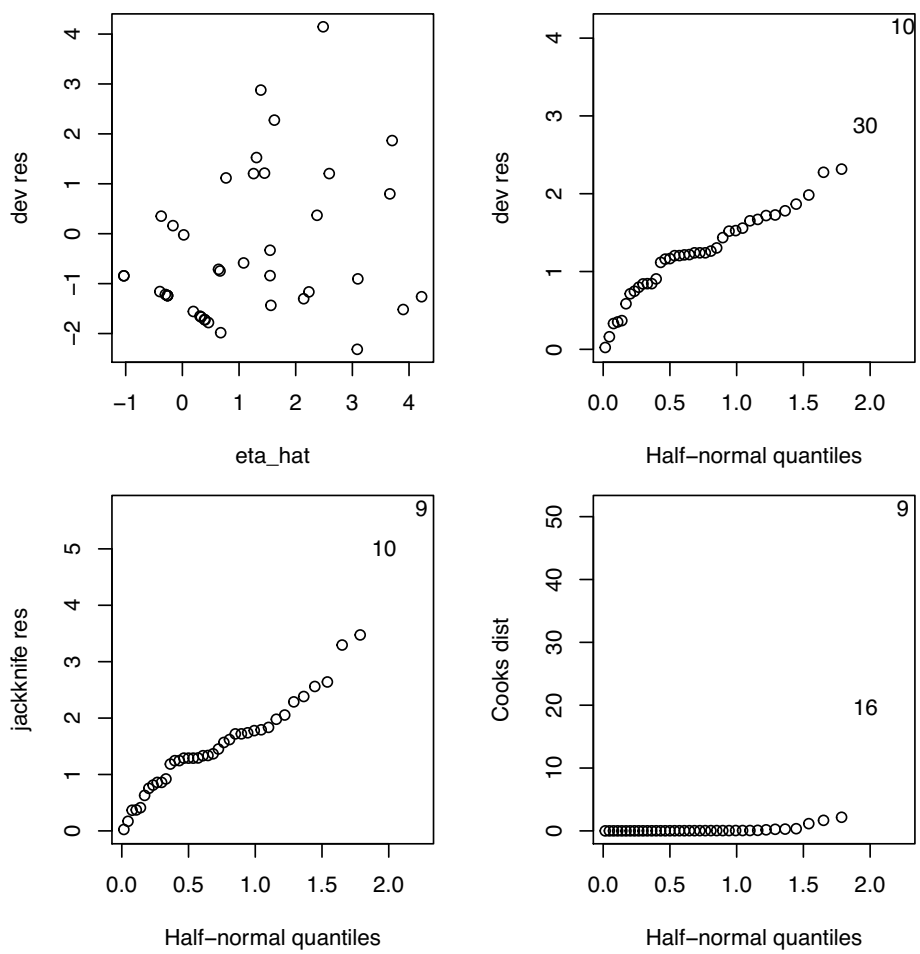



```
> modelA  <- glm(incidents ~ year + type + rootserv,
+                family=poisson(link="log"), ships)
> summary(modelA)

Call:                    1+4+2+1
glm(formula = incidents ~ year + type + rootserv, family = poisson(link = "log"),
    data = ships)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3728  -1.4400  -0.9767   0.6629   4.5219
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.167376   0.256858  -0.652 0.514640
year65       1.084977   0.157998   6.867 6.56e-12 ***
year70       1.626040   0.196274   8.285  < 2e-16 ***
year75       1.026427   0.261876   3.920 8.87e-05 ***
typeCD      -0.828441   0.215053  -3.852 0.000117 ***
rootserv     0.019527   0.001234  15.820  < 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: 730.25  on 39  degrees of freedom
Residual deviance: 105.42  on 34  degrees of freedom
AIC: 215.28

Number of Fisher Scoring iterations: 5

> modelB <- glm(incidents ~ year + type + rootserv + year:rootserv,
+               family=poisson(link="log"), ships)
> summary(modelB)

Call:
glm(formula = incidents ~ year + type + rootserv + year:rootserv,
    family = poisson(link = "log"), data = ships)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3162  -1.3362  -0.8438   0.4762   4.1455

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.193411   0.335567   0.576 0.564366
year65           0.703758   0.406551   1.731 0.083444 .
year70           1.304166   0.399825   3.262 0.001107 **
year75          -0.455966   0.532059  -0.857 0.391453
typeCD          -0.770174   0.222503  -3.461 0.000537 ***
rootserv         0.017483   0.001798   9.722  < 2e-16 ***
year65:rootserv  0.002173   0.002423   0.897 0.369942
year70:rootserv  0.001423   0.003372   0.422 0.672930
year75:rootserv  0.022332   0.006215   3.594 0.000326 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 730.253  on 39  degrees of freedom
Residual deviance:  90.276  on 31  degrees of freedom
AIC: 206.14

Number of Fisher Scoring iterations: 6

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

Analysis of Deviance Table

Model 1: incidents ~ year + type + rootserv
```

```
Model 2: incidents ~ year + type + rootserv + year:rootserv
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        34    105.418
2        31     90.276  3   15.142 0.001699 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> library(faraway) # for halfnorm function
> eta_hat <- predict(modelB, type="link")
> D_res <- residuals(modelB)
> J_res <- rstudent(modelB)
> Cooks <- cooks.distance(modelB)
> par(mfrow=c(2,2))
> par(mar=c(4,4,1,2))
> plot(eta_hat, D_res, ylab="dev res")
> halfnorm(D_res, ylab="dev res")
> halfnorm(J_res, ylab="jackknife res")
> halfnorm(Cooks, ylab="Cooks dist")
```




```
> (phi_hat <- sum(residuals(modelB, type="pearson")^2)/modelB$df.residual)
```

```
[1] 2.829891
```

```
> modelC  <- glm(incidents ~ type + year + rootserv,
+                family=quasipoisson(link="log"), ships)
> modelD <- glm(incidents ~ type + year + rootserv + year:rootserv,
+               family=quasipoisson(link="log"), ships)
> anova(modelC, modelD, test="F")
```

```
Analysis of Deviance Table

Model 1: incidents ~ type + year + rootserv
Model 2: incidents ~ type + year + rootserv + year:rootserv
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        34    105.418
2        31     90.276  3   15.142 1.7836 0.1708
```

(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) Write down the model fitted as modelB, including the fitted parameter values.

(d) Give an estimate for the expected number of incidents for a ship of type A constructed in 1970–1974, with 250 months service.

(e) Give two (quantitative) reasons why modelB may suffer from overdispersion.

(f) In two or three sentences, explain the difference between a Poisson and quasi-Poisson model.

(g) Give a std. error for the interaction between rootserv and level 75 of year, in the case where we allow for overdispersion.

(h) Allowing for overdispersion, do you perfer modelC or modelD, and why?

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

(j) If you were to choose one point to investigate as a potential outlier, which would you choose, and why?

If you removed this potential outlier from the analysis, would you then do about the overdispersion?

Question 4 (8 marks) 800 boys were classified according to socioeconomic status (S), whether the boy is a scout (B), and his juvenile delinquency status (D).

```
> boys <- read.csv("boys.csv")
> ftable(xtabs(y ~ S + B + D, boys))

         D  No Yes
S    B
High No     59   2
     Yes   196   8
Low  No    169  42
     Yes    43  11
Med  No    132  20
     Yes   104  14
```

We are interested in the dependence (if any) between the variables S, B, and D. Eight Poisson regression models were fitted, with different combinations of interactions. Here are the residual deviances

```
> deviance(glm(y ~ S + B + D, data=boys, family=poisson))

[1] 218.6622

> deviance(glm(y ~ S*B + D, data=boys, family=poisson))

[1] 36.41467

> deviance(glm(y ~ S + B*D, data=boys, family=poisson))

[1] 211.0496

> deviance(glm(y ~ S*D + B, data=boys, family=poisson))

[1] 182.4098

> deviance(glm(y ~ S*B + S*D, data=boys, family=poisson))

[1] 0.1623331

> deviance(glm(y ~ S*B + B*D, data=boys, family=poisson))

[1] 28.8021

> deviance(glm(y ~ S*D + B*D, data=boys, family=poisson))

[1] 174.7973

> deviance(glm(y ~ S*B + S*D + B*D, data=boys, family=poisson))

[1] 0.1542901
```

(a) What are the residual degrees of freedom (d.f.) for each model?

(b) Give an interpretation of each of the following three models. That is, what sort of dependence structure between the variables is being assumed?

(i) S + B + D (ii) S*B + D (iii) S*B + S*D

(c) Which dependencies are significant in this data (at the 95% level)? You will find the following chi-squared percentage points useful:

```
> qchisq(0.95, df=1:7)

[1]  3.841459  5.991465  7.814728  9.487729 11.070498 12.591587 14.067140
```

Question 5 (8 marks)

(a) Explain why the following two functions produce pseudo-random numbers with the same distribution.

```
> f1 <- function() {
+   U <- runif(1)
+   if (U < 0.1) {
+     return(0)
+   } else if (U < 0.3) {
+     return(1)
+   } else if (U < 0.6) {
+     return(2)
+   } else {
+     return(3)
+   }
+ }
> f2 <- function() {
+   U <- runif(1)
+   if (U < 0.4) {
+     return(3)
+   } else if (U < 0.7) {
+     return(2)
+   } else if (U < 0.9) {
+     return(1)
+   } else {
+     return(0)
+   }
+ }
```

(b) Sketch the cdf being simulated from above.

(c) What is the expected number of comparisons made by `f1` and `f2`? That is, on average how many times do we evaluate statements of the form `U < x`?

Which function will run faster?

Question 6 (10 marks)

(a) Suppose that $X|p \sim \text{binom}(m, p)$ and $p \sim \text{beta}(\alpha, \beta)$. X is said to have a beta-binom(m, α, β) distribution. What is $\mathbb{P}(X = x)$ for $x \in \{0, 1, \ldots, m\}$?

(b) Suppose we draw n items from a population of size N without replacement. Suppose K elements of the population are successes and the rest failures, and let the number of successes in our sample be $Y \in \{\max(0, n + K - N), \ldots, \min(n, K)\}$. Then Y is said to have a hypergeometric distribution, and

$$\mathbb{P}(Y = y) = \frac{\binom{K}{y}\binom{N-K}{n-y}}{\binom{N}{n}}.$$

Show that if K has a beta-binom(N, α, β) prior distribution, then the posterior distribution of K given y is a shifted beta-binomial, and obtain its (hyper-)parameters.

Question 7 (10 marks)

(a) Specify a Bayesian version of modelB from Question 3. Your model description should include a directed acyclic graph (DAG) showing the relationships between the stochastic and logical nodes, distributions for all the stochastic nodes, and descriptions of all the constants.

Use appropriate vague priors for all the parameters.

(b) How would you change your model to deal with overdispersion?

Question 8 (10 marks)

(a) State the Metropolis-Hastings algorithm for simulating from some distribution π on R^d.

(b) Prove that the M-H chain is reversible w.r.t. π.

(c) Explain in a few sentences what a burn-in period is, and why it might be necessary.

End of Exam—Total Available Marks = 80