THE UNIVERSITY OF MELBOURNE

Semester 2 Assessment, 2014

Department of Mathematics and Statistics

MATH30027 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:

- A single two-sided hand-written A4 sheet of notes.
- Scientific calculators. Graphical calculators are *not* allowed.

Instructions to Students

- You may remove this question paper at the conclusion of the examination
- You should attempt all questions. Marks for individual questions are shown.

Instructions to Invigilators

- Students may remove this question paper at the conclusion of the examination

Blank page (ignored in page numbering)

Question 1 (8 marks) The exponential density is $f(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

(a) Write f in the correct form for an exponential family. Identify the canonical parameter θ, dispersion parameter ϕ, and the functions a, b, and c.

(b) What are the canonical link and variance function for a GLM with a response following the exponential distribution?

(c) Identify a practical difficulty that may arise when using the canonical link in this case.

(d) When comparing nested models in this case, should we use an F test or a χ^2 test? Explain.

(e) Express the deviance in this case in terms of the responses y_i and the fitted values $\hat{y}_i$.

Question 2 (9 marks) Suppose that $Y_1, \ldots, Y_n$ are i.i.d. Poisson(λ); $\mathbb{P}(Y_i = y) = e^{-\lambda}\lambda^y/y!$.

(a) Derive the maximum likelihood estimator $\hat{\lambda}$ of λ.

(b) What is the Fisher information for λ? Hence give an asymptotic distribution for $\hat{\lambda}$.

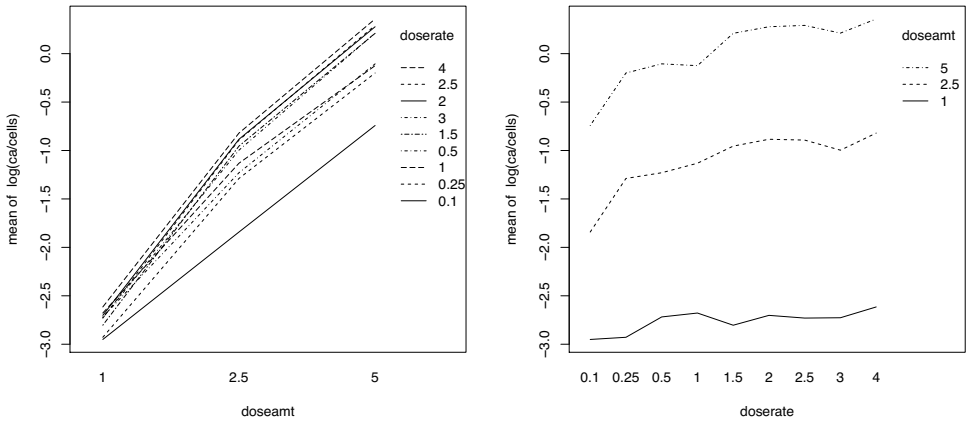
Suppose now that the Y_i have mean λ and variance $\phi\lambda$, for some $\phi > 0$.

(c) What is the quasi-(log)likelihood for λ?

(d) Derive the quasi-likelihood estimator for λ, and state an estimator for ϕ.

Question 3 (19 marks) In Purott and Reeder (1976), some data is presented from an experiment conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities (ca) observed. The observed counts ranged from 25 to 419. The number (cells), in hundreds of cells exposed in each run, differs. The dose amount (doseamt) and the rate (doserate) at which the dose is applied are the predictors of interest.

```
> library(faraway)
> data(dicentric)
> par(mfrow=c(1,2))
> with(dicentric, interaction.plot(doseamt, doserate, log(ca/cells), lty=c(1,2,4,5,6,1,2,4,5)))
> with(dicentric, interaction.plot(doserate, doseamt, log(ca/cells), lty=c(1,2,4)))
> par(mfrow=c(1,1))
```



Some R output from an analysis of these data follows. Use it to answer the questions that follow.

```
> model <- glm(ca ~ offset(log(cells)) + doserate*doseamt, family=poisson, data=dicentric)
> summary(model)

Call:
glm(formula = ca ~ offset(log(cells)) + doserate * doseamt, family = poisson,
    data = dicentric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7308  -2.2842  -0.6264   3.3487   5.8272

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.29994    0.06160 -53.567  < 2e-16 ***
doserate          0.06401    0.02922   2.191 0.028476 *
doseamt           0.61224    0.01707  35.862  < 2e-16 ***
doserate:doseamt  0.02715    0.00765   3.549 0.000387 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:  270.26  on 23  degrees of freedom
AIC: 453.67

Number of Fisher Scoring iterations: 4

> anova(model, test="Chi")

Analysis of Deviance Table

Model: poisson, link: log

Response: ca

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                 26     4753.0
doserate          1    231.3        25     4521.7 < 2.2e-16 ***
doseamt           1   4238.7        24      282.9 < 2.2e-16 ***
doserate:doseamt  1     12.7        23      270.3 0.0003681 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Questions:

(a) Specify the model fitted.

(b) What evidence do you have to support the linear assumptions in the model?

(c) The significance of the interaction term was tested twice, giving p-values of 0.000387 and 0.0003681. Define the test statistic used in each case, and the distributions used to obtain the

p-values. (You may use terms such as deviance and Fisher information without giving their exact formula.)

The results are very similar for this example, none-the-less we generally prefer one of these tests over the other. Which one?

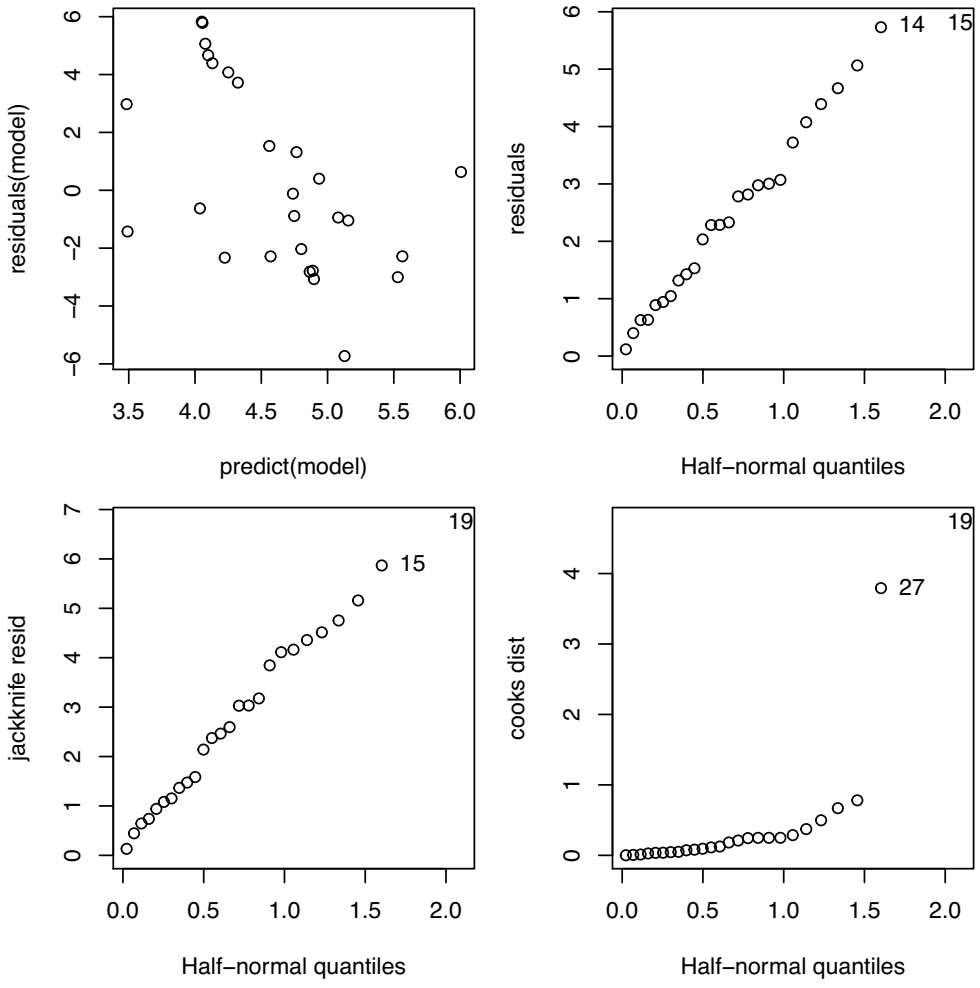
(d) What distribution will the residual deviance have (approximately)?

Does this type of distribution always apply for Poisson regressions? If not, when does it break down?

(e) Based on the residual deviance, do we reject or retain the hypothesis that the model is adequate (it explains the variation seen in the data)? Justify your answer.

Some further analysis and questions follow.

```
> par(mfrow=c(2,2),mar=c(4,4,1,1))
> plot(predict(model), residuals(model))
> halfnorm(residuals(model), ylab="residuals")
> halfnorm(rstudent(model), ylab="jackknife resid")
> halfnorm(cooks.distance(model), ylab="cooks dist")
> par(mfrow=c(1,1))
```




```
> (phi <- sum(residuals(model, type="pearson")^2)/23)
```

```
[1] 12.97226

> drop1(model, scale=phi, test="F")

Single term deletions

Model:
ca ~ offset(log(cells)) + doserate * doseamt

scale:  12.97226

                Df Deviance    AIC F value Pr(>F)
<none>               270.26 453.67
doserate:doseamt  1   282.95 452.64  1.0798 0.3096

> model2 <- glm(ca ~ offset(log(cells)) + doserate + doseamt, family=poisson, data=dicentric)
> (phi2 <- sum(residuals(model2, type="pearson")^2)/24)

[1] 12.72343

> summary(model2, dispersion=phi2)

Call:
glm(formula = ca ~ offset(log(cells)) + doserate + doseamt, family = poisson,
    data = dicentric)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.761  -1.696  -0.401   3.286   5.798

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.46115    0.15430 -22.432   <2e-16 ***
doserate     0.15501    0.04881   3.176   0.0015 **
doseamt      0.66230    0.03456  19.163   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 12.72343)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:  282.95  on 24  degrees of freedom
AIC: 464.35

Number of Fisher Scoring iterations: 4
```

(f) Are there any potential outliers?

(g) A problem with the model was unearthed. What, and how do you know?

(h) The significance of the interaction term has been tested again, and this time it is not significant. Describe the test statistic used and the distribution used to obtain the p-value.

(i) The model is refitted without the interaction term. Does the residual deviance tell us anything about the adequacy of this model?

Question 4 (5 marks) Radelet (1981) gave data on racial characteristics in homicide trials in Florida in 1976–77.

```
> data(death)
> death
```

```
    y penalty victim defend
1  19     yes      w      w
2 132      no      w      w
3   0     yes      b      w
4   9      no      b      w
5  11     yes      w      b
6  52      no      w      b
7   6     yes      b      b
8  97      no      b      b
```

The variables are

`penalty` did the defendent receive the death penalty?

`victim` was the victim black or white?

`defend` was the defendent black or white?

`y` count of the number of occurences.

We are interested in the dependence (if any) between the variables `penalty` (p), `victim` (v) and `defend` (d). Eight Poisson regression models were fitted, with different combinations of interactions. The residual deviances are

model	resid. deviance
p + v + d	137.9
p*v + d	131.7
p*d + v	137.7
p + v*d	8.13
p*d + v*d	7.91
p*v + v*d	1.88
p*v + p*d	131.5
p*v + p*d + v*d	0.70

(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) p + v + d (ii) p + v*d (iii) p*v + v*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=1:4)
```

```
[1] 3.841459 5.991465 7.814728 9.487729
```

Question 5 (8 marks) Consider the following WinBUGS code:

```
model {
  for (i in 1:n) {
    y[i] ~ dnorm(mu, 1)
  }
  mu ~ dnorm(10, 0.01)
}

# data
list(n = 6, y = c(6, 9, 8, 10, 7, 8))

# inits
list(mu = 10)
```

(a) Draw a directed acyclic graph (DAG) representing this model. Draw all stochastic and logical nodes but not constants.

(b) What are the prior, likelihood, and posterior distributions here?

Is the prior vague or informative? If it is vague, give an example of an informative prior, if it is informative, give an example of a vague prior.

(c) Suppose that you collected a sample of size 10,000 from the posterior. What values would you expect for the sample mean and sample standard deviation (approximately)?

Question 6 (4 marks)

(a) Show that if X has cdf F (assumed invertible) and $U \sim U(0,1)$ then $F^{-1}(U)$ has the same distribution as X.

(b) The Weibull(λ, k) distribution has density

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

Give an algorithm for simulating from this distribution.

You may assume that you have an infinite supply of i.i.d. $U(0,1)$ random variables.

Question 7 (15 marks) Suppose that we observe some variable y in m batches each of size n. Let y_{ij} be the j-th observation from batch i. Consider the following Bayesian model

$$\begin{aligned}
y_{ij} &\sim N(\mu_i, 1/\tau_w) \\
\mu_i &\sim N(0, 1/\tau_b) \\
\tau_w &\sim \text{gamma}(0.001, 0.001) \\
\sigma_w &\leftarrow 1/\sqrt{\tau_w} \\
\sigma_b &\sim U(0, 100) \\
\tau_b &\leftarrow 1/\sigma_b^2
\end{aligned}$$

(a) Draw the directed acyclic graph (DAG) corresponding to this model.

(b) Write down the kernel of the joint distribution of all the variables (observed and unobserved), except σ_w and τ_b. (Recall the gamma(α, β) density is proportional to $x^{\alpha-1}e^{-\beta x}$.)

(c) Is the prior on τ_w vague or informative?

What is the kernel of the marginal posterior of τ_w? Can you identify the distribution precisely?

(d) Put $ICC = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$. Suppose that rather than put a prior on σ_b^2 we wish to put a uniform prior on ICC. Modify the model to achieve this.

(e) Modify the original model so that rather than allowing the means to change from batch to batch, we allow the variances to change.

Question 8 (12 marks) Consider the following R coding of a Metropolis-Hastings algorithm:

```
p <- 0.5
r <- 2

nreps <- 1000000
X <- rep(NA, nreps)
X[1] <- 0
for (i in 2:nreps) {
  Y <- 0
  while (runif(1) > p) Y <- Y + 1
  alpha <- choose(Y+r-1, r-1)/choose(X[i-1]+r-1, r-1)
  if (runif(1) < alpha) {
    X[i] <- Y
  } else {
    X[i] <- X[i-1]
  }
}
```

Note that `choose(n, x)` returns $\binom{n}{x}$.

(a) What type of random variable is generated by the following two lines?

```
Y <- 0
while (runif(1) > p) Y <- Y + 1
```

(b) Is this a symmetric Metropolis-Hastings algorithm? Justify your answer.

(c) What is the stationary (and limiting) distribution of $X(n)$, the n-th element of the vector `X`?

(d) Suppose that you wish to estimate $\mathbb{P}(X \leq 2)$ under stationarity, using `mean(X <= 2)`

(i) Will this estimator be consistent? If so, what property of the Markov chain are you relying on?

(ii) If we discard `X[1:k]` for some `k`, could this improve our estimate? How might we choose `k` (a few of sentences will suffice to answer this)?

(iii) How can we gauge the accuracy of our estimate?

End of Exam—Total Available Marks = 80