University of Glasgow

Tuesday, 8th May 2018
2.00 pm - 4.00 pm

EXAMINATION FOR THE DEGREES OF M.A., M.SCI. AND B.SC. (SCIENCE)

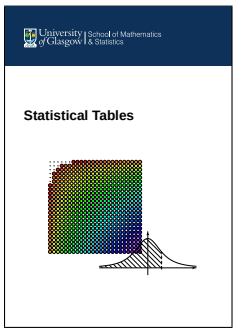
Statistics – Generalized Linear Models - Level M

This paper consists of 7 pages and contains 3 questions.
Candidates should attempt all questions.

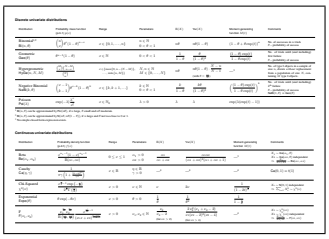
Question 1	22 marks
Question 2	25 marks
Question 3	33 marks
Total	80 marks

The following material is made available to you:

Statistical tables*



Probability formula sheet



"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

*Note for invigilators: will be delivered to the exam venue by the School

CONTINUED OVERLEAF/

1. (a) Consider the negative binomial distribution with probability mass function

$$f(y;\theta) = \binom{y+r-1}{r-1}\theta^r(1-\theta)^y, \qquad \theta \in (0,1),\ y = 0,1,2,\ldots,r$$

where $r > 0$ is assumed known. Show that this distribution is a member of the exponential family and identify its natural parameter. **[3 MARKS]**

(b) Let Y_i, $i = 1,\ldots,n$ be independent random variables from the above distribution with $\mu_i = E(Y_i)$. Consider two explanatory variables x_1 and x_2 that are each measured on the ith observation. Write down a generalized linear model for these data and suggest one suitable choice of link function, explaining why it is appropriate. Give the model in vector-matrix form, making sure to specify the response vector $\mathbf{y}$, the parameter vector $\boldsymbol{\beta}$ and the design matrix $\mathbf{X}$. **[5 MARKS]**

(c) Wildlife biologists want to model how many fish are caught by those fishing at a local park. They collect data on 250 groups visiting the park. Each group was questioned about how many fish they caught (`count`, an integer ranging from 0 to 149), how many children were in the group (`child`, an integer taking values 0, 1, 2 or 3), and whether or not they brought a campervan to the park (`camper`, taking values 1 for yes and 0 for no). The following R output shows the fit of a Poisson model (Model 1) and a negative binomial model (Model 2) to the data.

Output from Model 1:

```
Call:
 glm(formula = count ~ child + camper, family = poisson, data = fish)

Coefficients:
              Estimate Std. Error z value  Pr(>|z|)
  (Intercept)  0.91026    0.08119   11.21   <2e-16 ***
  child       -1.23476    0.08029  -15.38   <2e-16 ***
  camper       1.05267    0.08871   11.87   <2e-16 ***

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2958.4  on 249  degrees of freedom
Residual deviance: 2380.1  on 247  degrees of freedom
AIC: 2723.2
```

CONTINUED OVERLEAF/

Output from Model 2:

```
Call:
   glm.nb(formula = count ~ child + camper, data = fish,
          init.theta = 0.2552931119, link = log)

 Coefficients:
   Estimate Std. Error z value Pr(>|z|)
   (Intercept)   1.0727     0.2425   4.424 9.69e-06 ***
   child        -1.3753     0.1958  -7.025 2.14e-12 ***
   camper        0.9094     0.2836   3.206  0.00135 **

 (Dispersion parameter for Negative Binomial(0.2553) family taken to be 1)

 Null deviance: 258.93  on 249  degrees of freedom
 Residual deviance: 201.89  on 247  degrees of freedom
 AIC: 887.42

 Theta:  0.2553
 Std. Err.:  0.0329

 2 x log-likelihood:  -879.4210
```

Explain what is meant by overdispersion, and how it may arise when analysing count data. What is the role of the negative binomial distribution in modelling such data? Is there evidence of overdispersion in the above analyses?

[6 MARKS]

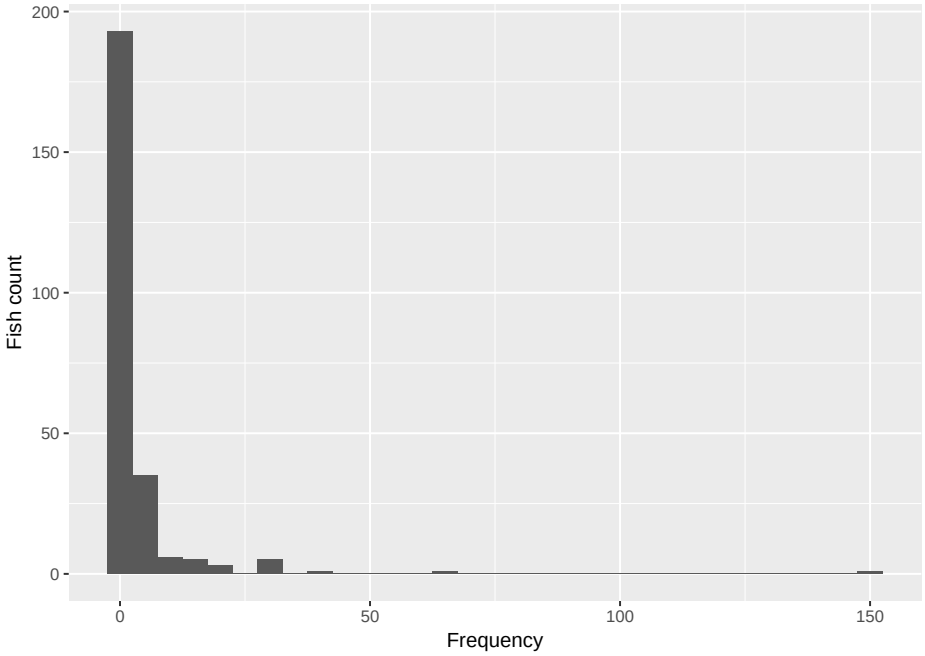
(d) Two groups arrive at the park with campervans, one group with children and one without children. Based on the output of Model 2, is the group with children expected to catch more or fewer fish than the group without children? Justify your answer. **[2 MARKS]**

(e) Two groups arrive at the park without children, one in a campervan, the other in another vehicle. Based on the Model 2 output, which group is expected to catch more fish? Justify your answer. **[2 MARKS]**

(f) Based on Model 2, what is the estimated count of fish caught by a group with a campervan and two children? **[2 MARKS]**

CONTINUED OVERLEAF/

(g) Examine the following histogram of the count of fish caught and suggest an alternative model that could be used to analyse these data.



[2 MARKS]

CONTINUED OVERLEAF/

2. A psychologist conducted a study to examine the nature of the relationship between an employee's emotional stability x and the employee's ability to perform in a task group Y. Emotional stability was measured by a written test for which the higher the score, the greater the emotional stability. Ability to perform in a task group ($Y = 1$ if able, $Y = 0$ if unable) was evaluated by the employee's supervisor.

   (a) Write down the generalized linear model being fit by the R command below:

```
fit <- glm(perform ~ stability, family = binomial)
```

   [5 MARKS]

   The following is partial output from fitting the model in R.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.308925   4.376997  -2.355   0.0185 *
stability     0.018920   0.007877   2.402   0.0163 *

    Null deviance: 37.393  on 26  degrees of freedom
Residual deviance: 29.242  on 25  degrees of freedom
AIC: 33.242
```

   (b) The Hosmer-Lemeshow test statistic for the above model is $X^2_{HL} = 4.8$ on 8 degrees of freedom. State the null hypothesis for this test and give your conclusions based on the value of the test statistic. **[4 MARKS]**

   (c) Assuming that the above model is adequate, test the hypothesis that the stability coefficient is equal to zero. **[2 MARKS]**

   (d) Interpret the stability coefficient by giving a relevant point estimate and an approximate 95% confidence interval for the odds of being able to perform in the task group. **[6 MARKS]**

   (e) What is the estimated probability that an employee with an emotional stability test score of 500 will be able to perform in a task group? **[3 MARKS]**

   (f) What is the emotional stability test score for which there is an estimated 70% probability of being able to perform in a task group? **[3 MARKS]**

   (g) Suppose that the ability of the employee to perform in a task group was rated on a scale from 1 to 5, with 5 for extremely able and 1 for not able at all. What model would you use to fit to the data in this case? **[2 MARKS]**

CONTINUED OVERLEAF/

3. In a physical process, photons are emitted with intensity proportional to the density of the image. At the ith point, the number of detected photons, Y_i, has a Poisson distribution with mean μ_i which is a linear function of the densities β of different points of the image.

   In a particular experiment, we have two points where the photons are emitted with the corresponding image densities β_1 and β_2, and we observe the number of emitted photons y_i at six points, $i = 1, \ldots, 6$. The connection between the mean μ_i at the ith point and the image densities β_1 and β_2 is given by

$$\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2}, \qquad i = 1, \ldots, 6. \tag{1}$$

   The number of observed photons y_i and the corresponding coefficients x_{i1} and x_{i2} at the ith point are shown in the table below.

Point	y_i	x_{i1}	x_{i2}
1	10	0.78	0.16
2	12	0.72	0.51
3	12	0.42	0.88
4	9	0.16	0.83
5	2	0.04	0.43
6	1	0.01	0.12

   (a) State the model as a generalized linear model. Would the canonical link function be appropriate here? What constraints should the estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and the fitted values satisfy? **[8 MARKS]**

   (b) Results of an analysis of the data set using the Poisson model with the mean given by (1) are given below (Model 1).

   Model 1 output:

```
> glm(y ~x1+x2-1, data=photons, family=poisson(link="identity"))

Call:  glm(formula = y ~ x1 + x2 - 1, family = poisson(link = "identity"),
    data = photons)

Coefficients:
    x1      x2
11.867   7.073

Degrees of Freedom: 4 Residual
Residual Deviance: 1.033  AIC: 26.53

> fitted(glm(y ~x1+x2-1, data=photons, family=poisson(link="identity")))
         1          2          3          4          5          6
         ? 12.1514735 11.2081319  7.7690030  3.5158977  0.9673816
```

CONTINUED OVERLEAF/

Calculate the fitted value $\hat{y}_1$ that has been omitted from the R output above. Do the estimates of β_1 and β_2 and the fitted values satisfy the required constraints? **[4 MARKS]**

(c) Suppose that we want to test whether to add the intercept. For the model with the intercept

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \qquad i = 1, \ldots, 6, \tag{2}$$

the generalized linear model fit is given below (Model 2).

Model 2 output:

```
> glm(y ~x1+x2, data=photons, family=poisson(link="identity"))

Call:  glm(formula = y ~ x1 + x2, family = poisson(link = "identity"),
    data = photons)

Coefficients:
(Intercept)           x1           x2
    -0.2366      11.9139       7.5233

Degrees of Freedom: 3 Residual
Null Deviance:     20.26
Residual Deviance: 0.9784  AIC: 28.48
```

Using the output from both models, test the hypothesis that $\beta_0 = 0$, i.e. whether the model without the intercept is appropriate. **[5 MARKS]**

(d) Explain why considering raw residuals would be inappropriate and name an alternative residual that can be used for diagnostic checks. **[3 MARKS]**

(e) Write down the log-likelihood for model (1) and obtain the likelihood equations for β_1 and β_2. Name the numerical algorithm that is used to solve them. **[5 MARKS]**

(f) Derive an expression for the deviance of model (1) and use its numerical value from the output of Model 1 to check the goodness of fit of the model. Do you have any concerns about using the deviance in this way? **[8 MARKS]**

END OF QUESTION PAPER.