University of Glasgow

Wednesday, 20th May 2020
09:30 - 13:30 BST

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

Statistics – Generalised Linear Models - Level M - STATS5019

1. **You are about to sit an online assessment.** You have a set period of four hours to complete this exam. It is not expected that you will use the whole of this time. The unadjusted exam duration is *half* this time. The additional time allows for downloading the paper, uploading your answers and for any adjustments you might normally receive if you are registered with the University's Disability Service.

2. **Enlarging the text.** In case you should need to enlarge the text of a `.pdf` document: use your PDF viewer software to change to the desired magnification/zoom level.

3. **Advice on the contents of the exam and technical support.** In case you should have questions about the contents of this paper or you require technical assistance, please contact our virtual invigilation team at the University of Glasgow Helpdesk `https://www.gla.ac.uk/help`. An academic member of staff will be available to answer questions about the examination, as they would normally, at the beginning of the exam. Technical support will be available 24 hours per day. To ensure timely responses and to ensure that all students receive the same information, you should not contact academic staff directly but always use the Helpdesk.

4. **Submitting your answers.** Acceptable file types for submitting typed documents are: `.doc/.docx`, `.rtf`, `.pdf`, `.xls/.xlsx`. Acceptable file types for submitting high resolution images are: `.jpg`, `.png`, `.tif`, `.pdf`. In case you are unable to upload your answers to Moodle, you may email them to:

   `maths-stats-exams@glasgow.ac.uk`.

   Keep the original paper copies.

CONTINUED OVERLEAF/

5. **Recommended approach to writing your exam and uploading your solutions.**

   - Write your solutions on paper.
   - Number the pages sequentially and write your matriculation number and the course code of the exam in the top left of every page.
   - Photograph or scan the pages, one by one.
   - Check that the images are clear enough for us to read.
   - Combine pages in order, if possible. For example, use a smartphone app like Microsoft Office Lens or Adobe Scan, which can do this automatically, or add photos to a Microsoft Word document (part of Office 365).
   - Ideally, save as a `.pdf` file on your device.
   - Name the file with the convention: `XXXXXYYYY-ZZZZZZZz-N.ext`. Here `XXXXX` = `"STATS"` or `"MATHS"`; `YYYY` = course code number; `ZZZZZZZz` = matriculation number including final letter; `N` = number of upload; `.ext` = file extension. For example: `STATS5999-9876543s-1.pdf`. Do **not** include your name in the file name.
   - Upload that file (or files) to Moodle inside the exam Assignment.
   - Note that there is an upload limit of 100MB per file. If necessary, you can break up the solution into multiple files. However, Moodle does not allow more than 20 files.
   - You can word-process your answers if you wish (although it is probably less efficient). Then you would just upload that document (ideally saved as `.pdf`).

6. **Declaration of Academic Integrity.** Your answers must entirely be your own work. During the period of time that this exam is active, you must not for any reason communicate or collude with other students taking this exam. Note that your exam papers may be processed through Turnitin for plagiarism checking. We may also conduct a further oral examination to check your knowledge and establish that the exam answers are your own. This declaration incorporates the University's Declaration of Originality which applies to all academic work.

7. **Declaring that the work is your own.** In order to view this exam paper, you must have checked the box in Moodle to agree to both this declaration and the University's Declaration of Originality.

CONTINUED OVERLEAF/

This paper consists of 8 pages and contains 4 questions.

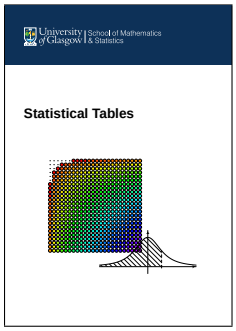
Candidates should attempt *ALL* questions.

This table shows an indicative number of marks for each question:

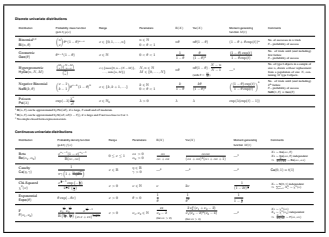
Question 1	20 marks
Question 2	20 marks
Question 3	20 marks
Question 4	20 marks
Total	80 marks

The following material is available to you on Moodle:

Statistical tables



Probability formula sheet



Any electronic calculator may be used.

CONTINUED OVERLEAF/

1. (a) Describe two ways in which generalised linear models extend the normal linear model. **[4 MARKS]**

   (b) Suppose that Y is a random variable with a distribution $f(y;\theta)$ that is a member of the exponential family of distributions with probability density (or mass) function

   $$f(y;\theta) = \exp[yb(\theta) + c(\theta) + d(y)].$$

   Define the *score function* $U(\theta, Y)$ for a single observation. Derive an expression for the mean of Y using a property of the score function. **[3 MARKS]**

   (c) Show that the exponential distribution with probability density function

   $$f(y;\theta) = \theta e^{-\theta y}, \qquad y > 0 \text{ and } \theta > 0,$$

   is a member of the exponential family of distributions and that it has a probability density function of the form given in part (b) above. **[3 MARKS]**

   (d) Using the expression for the expected value of Y obtained in part (b), show that $E(Y) = \dfrac{1}{\theta}$ for the exponential distribution. Deduce the canonical link function which would be appropriate for use in a generalised linear model. **[4 MARKS]**

   (e) Suppose that $y_1, \ldots, y_n$ are independent observations each from an exponential distribution which depends on the value of an explanatory variable x through the model $\frac{1}{\mu_i} = \beta_0 + \beta_1 x_i$ where $\mu_i = E(Y_i)$. Derive expressions for the score vector and information matrix and describe how they are both used to obtain the maximum likelihood estimates of the parameters β_0 and β_1.

   [6 MARKS]

2. Data are available on the number of fatalities in traffic accidents in the US for seven consecutive years in the period 1982-1988. The `Fatalities` dataset contains 334 observations and 34 variables. The outcome of interest is `fatal`, the number of fatalities that occurred in each state for each year of available data. The explanatory variables considered in this analysis are `year`, taking values from 1982 to 1988, `state` (48 in total, listed alphabetically from Alabama to Wyoming) and `beertax`, the percentage of tax on the price of a case of beer. The offset is `log(milestot)`, where `milestot` is the total vehicle miles (in millions) travelled in each state for each year of the study.

   Fitting a Poisson regression model with the log link to these data gave the following results:

```
> glm(fatal ~ year+state+beertax, offset = log(milestot),
          family = poisson, data=Fatalities)
Coefficients:
```

CONTINUED OVERLEAF/

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.985399   0.063303 -47.161  < 2e-16 ***
year1983    -0.068848   0.006826 -10.086  < 2e-16 ***
year1984    -0.085171   0.006773 -12.575  < 2e-16 ***
year1985    -0.119757   0.006816 -17.570  < 2e-16 ***
year1986    -0.113580   0.006798 -16.709  < 2e-16 ***
year1987    -0.152146   0.006924 -21.973  < 2e-16 ***
year1988    -0.196689   0.007122 -27.615  < 2e-16 ***
stateaz     -0.167350   0.052063  -3.214 0.001307 **
... [output omitted]
statewy     -0.387392   0.067346  -5.752 8.81e-09 ***
beertax     -0.292140   0.037654  -7.759 8.58e-15 ***

    Null deviance: 12293  on 335  degrees of freedom
Residual deviance:  1734  on 281  degrees of freedom
AIC: 4621.3
```

(a) Explain the role of the offset term in the model. **[2 MARKS]**

(b) Based on the output, what is the general time trend for the fatality rate? Quantify this by referring to the rate ratio which compares the expected fatality rate in 1988 to the rate in 1982. **[3 MARKS]**

(c) Interpret the `beertax` coefficient in terms of an appropriate rate ratio. Your answer should include both a point estimate and a confidence interval.

[3 MARKS]

(d) Explain what is meant by overdispersion and describe how it might arise in the above model. **[4 MARKS]**

(e) Describe two ways in which you could check for evidence of overdispersion in the above model.

[4 MARKS]

(f) List two ways in which you could further develop the model fitted above.

[4 MARKS]

CONTINUED OVERLEAF/

3. Six different doses of a drug aimed to decrease the cholesterol level were administered to patients with high cholesterol levels, and the outcome was recorded as 1 (success) if the patient's cholesterol level returned to normal after a month of treatment, and 0 (failure) otherwise. We are interested in estimating the probability of success p_i given dose x_i, for $i = 1, \ldots, 6$.

   The numbers of successes y_i (out of a total of n_i patients in the ith dose group) are given below.

Group	y_i	n_i	x_i
1	7	10	0.8
2	5	9	1.1
3	7	10	1.4
4	6	10	2.3
5	5	9	2.5
6	2	10	4.1

   (a) Show that the probit model for the probability of success $p_i = \Phi(\beta_0 + \beta_1 x_i)$ is equivalent to

$$p_i = \Pr(Z < x_i) \text{ if } \beta_1 > 0,$$
$$p_i = \Pr(Z > x_i) \text{ if } \beta_1 < 0,$$

   where $Z \sim N(-\beta_0/\beta_1; 1/\beta_1^2)$.

   Explain why the model with relationship $p_i = \Phi(\beta_0 + \beta_1 x_i)$ may be preferable to a model with relationship $p_i = \beta_0 + \beta_1 x_i$. **[6 MARKS]**

   (b) An analysis of the data in R using the **logit** model $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ gave the following results:

```
> glm(cbind(y,n-y)~x , family=binomial(logit))

Coefficients:
(Intercept)            x
     1.4087      -0.5855

Degrees of Freedom: 5 Total (i.e. Null);  4 Residual
Null Deviance:     7.15
Residual Deviance: 1.524  AIC: 21.35
```

   Using these estimates, calculate the fitted number of successes for group 2. **[3 MARKS]**

CONTINUED OVERLEAF/

(c) Interpret the estimated coefficient $\hat{\beta}_1$ from the logit model in terms of the odds of a success. **[2 MARKS]**

(d) Explain why considering raw residuals to check the model fit of the logit model would be inappropriate, and name two alternative types of residuals that would be appropriate to use for such a model. **[2 MARKS]**

(e) Using the R output in part (b),

i. test the goodness of fit of the model with $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$, stating a condition that the fitted values should satisfy for such a goodness-of-fit test to be valid; **[4 MARKS]**

ii. assuming that the model $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ is adequate, test the hypothesis that $\beta_1 = 0$. **[3 MARKS]**

4. A bank collected data on 1000 loan applicants. The outcome variable of interest (Y) takes the value 1 if the loan outcome was positive, and 0 otherwise. 300 loan applications resulted in positive outcomes, with the remaining 700 applications rejected. Potential explanatory variables include:

- C: the applicant's credit history, a categorical variable with five levels coded as
  - 0 : no credits taken
  - 1 : all credits at this bank paid back duly
  - 2 : existing credits paid back duly till now
  - 3 : delay in paying off in the past
  - 4 : critical account;
- A: the loan amount (in pounds);
- D: the loan duration (in months).

CONTINUED OVERLEAF/

The following table summarises the results of fitting binary logistic regressions to the above data, treating credit history, loan amount and duration as explanatory variables.

Variables in model	Deviance	Number of parameters in model
Null		
C	1199.06	
A	1090.39	
D	1177.11	
C + A	1070.47	
C + D	1051.90	
A + D	1176.55	
C + A + D	1051.19	

(a) Show that the maximised log-likelihood of the null model is given by

$$y \log y + (n - y) \log(n - y) - n \log n.$$

[6 MARKS]

(b) Calculate the null deviance value that should go in the first row of the above table.
[2 MARKS]

(c) Using a generalised likelihood ratio test procedure, select the simplest model that best fits the data from those listed in the table. **[5 MARKS]**

(d) Write down the logistic regression equation for the model you have selected, clearly defining any additional notation that you use. Explain how the bank can use this logistic regression model to evaluate the credit worthiness of a loan applicant.
[4 MARKS]

(e) Suppose that an additional 1000 observations not already included in fitting the model are available for use as a test set. List three measures of predictive performance that you would apply to the test set to assess the model.
[3 MARKS]

END OF QUESTION PAPER.