Instituto Superior de Economia e Gestão - UL

MSc Actuarial Science

Generalised Linear Models

9th June 2015

Time allowed: two hours and a half

Instructions

- This paper contains 3 questions, and comprises 5 pages including the title page.
- Enter all the requested details on the cover sheet.
- You have ten minutes reading time. You must not start writing your answers until instructed to do so.
- Number the pages of the paper where you are going to write your answers.
- Attempt all 3 questions.
- Begin your answer to each of the 3 questions on a new page.
- Marks are shown in brackets. Total marks: 200.
- Show calculations where appropriate.
- An approved calculator may be used.
- The Formulae and Tables for Actuarial Examinations (the 2002 edition) may be used.

1 - Suppose that a random variable Y_i is a member of the exponential family, i.e. its density function is given by:

$$f_Y(y_i;\theta_i,\varphi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i,\varphi)\right]$$

a) [15] Show that if $Y_i \sim$ *Exponential* of mean μ_i, then Y_i is a member of the exponential family. Identify the natural and scale parameters and functions $a(\varphi)$, $b(\theta)$ and $c(y,\varphi)$.

b) [15] Using the properties of the exponential family of distributions, derive an expression for the mean and for the variance of Y as functions of the natural parameter θ_i.

In the context of a generalised linear model, consider now the following linear predictor: $\eta_i = \alpha + \beta x_i + \gamma x_i^2$, where x_i is a continuous variable. Let $g(.)$ be a link function such that $g(E(Y_i)) = \eta_i$.

c) [15] Explain what is the "canonical" link function, and identify it for the Exponential distribution. Express the mean of Y as a function of the proposed linear predictor for the canonical link.

d) [10] Maximum likelihood estimation, performed on a random sample of 120 observations, gave the following results: $\widehat{\alpha} = 0.2381$ with $se(\widehat{\alpha}) = 0.2647$, $\widehat{\beta} = 0.0116$ with $se(\widehat{\beta}) = 0.0042$, and $\widehat{\gamma} = 0.0229$ with $se(\widehat{\gamma}) = 0.0197$. A researcher argued that variable x does not have a quadratic effect. Computing a suitable confidence interval, comment on the researcher's statement.

2 - In order to study the effects of alcohol on driving, information on a random sample of 9772 car accidents was collected, having recorded the following variables: driver's gender (Male/Female), driver's blood alcohol concentration (High/Low), and whether the driver was responsible for the accident (Guilty/Not Guilty), having obtained the results below:

		Males		**Females**		
		High C.	**Low C.**	**High C.**	**Low C.**	**Total**
Driver	**Guilty**	353	3357	243	2813	6766
	Not Guilty	20	1507	65	1414	3006
Total no. of accidents		373	4864	308	4227	9772

a) [15] Within the exponential family of distributions, state what would be an appropriate model for the probability of being guilty as a function of the alcohol concentration and gender. Propose a possible link function that you would suggest to use with the chosen distribution. Justify.

Using the above data, the following variables were defined: `gender` (gender); `alcohol` (blood alcohol concentration); `guilty` (number of accidents with guilty drivers) and `accidents` (total number of accidents). Maximum likelihood estimation gave the following results:

```
Call:  glm(formula = guilty/accidents ~ alchool + gender, family = binomial,
          weights = accidents)

Coefficients:
               Estimate  Std.Error  z value  Pr(>|z|)
 (Intercept)    0.82370    0.03087   26.681   < 2e-16
 alchoolHigh    1.19904    0.11814   10.149   < 2e-16
 genderFemale  -0.16082    0.04417   -3.641  0.000271
```

b) [15] According to the above model, derive the expression of the probability of being responsible for an accident as a function of the linear predictor η. Estimate such probability for a male with high blood alcohol concentration.

c) [15] A second model was proposed, with the same family and link function, where the linear predictor is function of blood alcohol concentration only. Justifying through an appropriate test, say whether you would prefer this model to the previous one.

d) [10] Would you consider appropriate using the Poisson distribution for the modelling of the rate of accidents with guilty driver? Justify.

3 - A researcher studying peptic ulcer in the UK collected data for three cities, by gender and blood type, having obtained the following number of cases (per million person-years):

gender	blood	city	cases
"Male"	"A"	"London"	579
"Female"	"A"	"London"	421
"Male"	"0"	"London"	911
"Female"	"0"	"London"	457
"Male"	"A"	"Manchester"	377
"Female"	"A"	"Manchester"	246
"Male"	"0"	"Manchester"	526
"Female"	"0"	"Manchester"	291
"Male"	"A"	"Newcastle"	453
"Female"	"A"	"Newcastle"	361
"Male"	"0"	"Newcastle"	459
"Female"	"0"	"Newcastle"	396

The researcher would like to model the incidence of peptic ulcer as function of the above three factors, and decides to estimate a generalised linear model.

Maximum likelihood estimation results are shown in the `R` output below. The dependent variable `cases` is the number of individuals with peptic ulcer (per million person-years), and the explanatory variables are `gender` (two-level factor, Male=1, Female=2), `blood` (two-level factor, blood group A or 0), and `city` (three-level factor, levels as in table).

Model 1

```
Call:  glm(formula = cases ~ gender + city + blood, family = poisson)

Coefficients:
                 Estimate  Std.  Error  z value  Pr(>|z|)
 (Intercept)      6.45489    0.02775    232.611   0.0000
 gender Female   -0.41979    0.02762    -15.198   0.0000
 city Manchester -0.49740    0.03342    -14.884   0.0000
 city Newcastle  -0.34982    0.03196    -10.946   0.0000
 blood O          0.22109    0.02719      8.131   0.0000

(Dispersion parameter for poisson family taken to be 1)

Null deviance:  623.575 on 11 degrees of freedom
Residual deviance:  72.055 on 7 degrees of freedom
```

a) [15] Describe the proposed model, specifying the distribution, linear predictor and link function. Specify a parametrised expression for the linear predictor.

b) [10] Give the interpretation of the intercept parameter, and of the coefficient denoted by `"blood O"`.

c) [15] A student argues that what matters in terms of location is only whether it is in London or not. How would you express such hypothesis in terms of the parameters of Model 1? Test the hypothesis, knowing that the covariance between the estimates of the coefficients of `"city Manchester"` and `"city Newcastle"` is equal to 0.000422297. What is your conclusion?

Keeping the same family and link function as in Model 1, the researcher estimates two additional models, called Model 2 and 3 respectively, in order to consider alternative specifications for the linear predictor. The full results are not shown here, but are summarised in the following deviance table.

Analysis of Deviance Table

```
anova(m1, m2, m3, test = "Chi")

Model 1:  cases   gender + city + blood
Model 2:  cases   gender * blood + city
Model 3:  cases   gender * (city + blood)

    Resid.Df  Resid.Dev  Df  Deviance   Pr(>Chi)
 1      7       72.055
 2      6       60.357    1   11.698   0.0006258
 3      4       27.988    2   32.368    9.36e-08
```

d) [15] Describe in detail Model 2 and 3 above, specifying a parametrised expression for the linear predictor and justifying the degrees of freedom.

e) [10] Based on the results of the deviance table, would you consider adding interactions to the original model? If yes, which of the alternative models would you choose? Justify specifying the appropriate tests.

f) [15] What is the objective of the estimation of Model 4 below? Explain in detail the difference between Model 4 and Model 3 presented earlier, and justify the application of Model 4 to the initial data.

g) [10] According to results shown in the last two rows of the deviance table associated to Model 4, does you conclusion in e) change now? Justify.

Model 4

```
Call:  glm(formula = cases ~ gender * (city + blood),
                    family = quasi(link = log, variance = "mu"))

Coefficients:
                              Estimate  Std.Error  t value  Pr(>|t|)
 (Intercept)                   6.45397    0.08705   74.143  1.98e-07
 gender Female                -0.42436    0.13850   -3.064    0.0375
 city Manchester              -0.50081    0.11180   -4.480    0.0110
 city Newcastle               -0.49089    0.11145   -4.405    0.0116
 blood O                       0.29687    0.09324    3.184    0.0334
 genderFemale:cityManchester   0.00916    0.18327    0.050    0.9625
 genderFemale:cityNewcastle    0.34261    0.17236    1.988    0.1178
 genderFemale:bloodO          -0.18995    0.14721   -1.290    0.2665

(Dispersion parameter for quasi family taken to be 7.027038)

Null deviance:  623.575 on 11 degrees of freedom
Residual deviance:  27.988 on 4 degrees of freedom
```

Analysis of Deviance Table

```
Model:  quasi, link:  log

Response:  cases

Terms added sequentially (first to last)
               Df  Deviance Resid.  Df  Resid.  Dev        F    Pr(>F)
 NULL                               11     623.58
 gender         1      236.079      10     387.50     33.5958  0.004405
 city           2      248.919       8     138.58     17.7115  0.010295
 blood          1       66.523       7      72.05      9.4667  0.037047
 gender:city    2       32.368       5      39.69      2.3031  0.216018
 gender:blood   1       11.698       4      27.99      1.6647  0.266508
```