Instituto Superior de Economia e Gestão - UL

MSc Actuarial Science

Generalised Linear Models

9th June 2015

Solution

1 -

a) [15] The density of an *Exponential* variable of mean μ_i can be re-written as:

$$f(y_i) = exp\left[-\frac{y}{\mu_i} - \log \mu_i\right]$$

which is in the form of an exponential family, where:

- the natural parameter is $\theta = -1/\mu$
- the scale parameter is $\varphi = 1$
- $a(\varphi) = \varphi = 1$
- $b(\varphi) = -\log(-\theta)$
- $c(y, \varphi) = 0$

b) [15] From the properties of the exponential family of distributions, we know that $E(Y) = b'(\theta)$, and $\text{Var}(Y) = a(\varphi)\ b''(\theta)$. Since $a(\varphi) = \varphi = 1$ and $b(\theta) = -\log(-\theta)$, then

- $b'(\theta) = -1/\theta$
- $b''(\theta) = 1/\theta^2$

It follows that $E(Y) = -1/\theta$ and $\text{Var}(Y) = 1 \times 1/\theta^2 = 1/\theta^2$.

c) [15] The canonical link function is such that the linear predictor equals the natural parameter, i.e. $\eta = \theta$, so that $g(E(Y)) = \theta$.

We showed in a) that the natural parameter is $\theta = -1/\mu$, and in b) that the expected value is $E(Y) = \mu$. Substituting into the above expression we obtain $g(\mu) = -1/\mu$, i.e. the canonical link function is the reciprocal. It follows that for the canonical link

$$E(Y_i) = -\frac{1}{\eta_i} = -\frac{1}{\alpha + \beta x_i + \gamma x_i^2}$$

d) [10] The hypothesis that that variable x does not have a quadratic effect corresponds to the hypothesis that $\gamma = 0$. In order to test it against a two-sided alternative, we can construct a confidence interval for γ and verify whether it contains the null value.

An approximate 95% c.i. is given by $\hat{\gamma} \pm 1.96 \times se(\hat{\gamma}) = 0.0229 \pm 1.96 \times 0.0197 = 0.0229 \pm 0.038612 = (-0.015712, 0.061512)$. Since the null value zero is contained in the interval, we can conclude that parameter γ is not significantly different from zero at a 5% level, and therefore the researcher is right.

2 -

a) [15] Since the response variable Y is a binary variable (say $Y = 1$ if the individual is guilty, $Y = 0$ otherwise), the appropriate distribution to choose within the exponential family is the Binomial, $Bin(n, \mu)$. The Binomial distribution in this case should be applied to the rates "no. guilty/no. accidents" considering the number of accidents n as weights.

Any link function such that $\mu = E(Y) = \Pr(Y = 1)$ is bounded between 0 and 1 is appropriate. For instance, we could choose the canonical link function for the Binomial, that is the logit:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

b) [15] The linear predictor η is linked to the probability of being responsible for an accident through the link function, since $\eta = g(\mu)$ and for the Binomial $\mu = \Pr(Y = 1)$. We obtain that $\Pr(Y = 1) = g^{-1}(\eta)$.

According to the above model, since no link function is specified, the software assumes the canonical link, i.e. the logit. Inverting the expression for the logit function, the required probability as a function of the linear predictor is given by:

$$\Pr(\text{being guilty}) = \frac{e^{\eta}}{1+e^{\eta}}$$

The estimated linear predictor for a male with high blood concentration is given by $\hat{\eta} = 0.82370 + 1.19904 = 2.02274$. Therefore the required estimated probability is:

$$\hat{p} = \frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} = \frac{e^{2.02274}}{1+e^{2.02274}} = \frac{7.5590}{8.5590} = 0.8832$$

c) [15] The alternative model would be preferred to the first one if the probability of being responsible for an accident does not depend on gender, i.e. if the coefficient denoted by `"genderFemale"` (say β_F) can be considered statistically equal to zero.

We can test the hypothesis using the test statistic $Z = \widehat{\beta}_F / se(\widehat{\beta}_F) \overset{a}{\sim} N(0,1)$. The observed value, from the output above, is given by:

$$z_{\text{obs}} = \frac{-0.16082}{0.04417} = -3.641$$

whose absolute value largely exceeds the critical value at 5% (1.96). Therefore, we reject the null hypothesis and conclude that the probability of being responsible for an accident depends on gender, so the first model is better. The observation of the p-vaule=0.000271 from the output would lead to the same conclusion.

d) [10] The Poisson distribution can be used to model rates as a large-sample approximation to the Binomial model in case of a "rare event", that is a binomial distribution $Bin(n, \theta)$ can be well approximated by a $Poisson(\lambda)$, with $\lambda = n\theta$ in the case of large n and small θ (usually less that 0.1).

The observation of our data set shows that, depsite the sample size being large, the event of being guilty cannot be considered "rare", since we observe that about 70% of the accidents have guilty driver. Therefore, in this case, altough we do not know the actual value of θ, we can infer from the sample evidence that it is larger than 0.1, and conclude that the Poisson approximation would not be appropriate.

3 -

a) [15] In the `R` output for Model 1 no link function is specified, so the software uses by default the canonical link for the specified family. In this case (Poisson), the canonical link is the logarithm, i.e. $\eta = g(\mu) = \log(\mu)$.

In the presented model, the linear predictor is a function of three factors (gender, blood type and city) without interactions, so that a parametrised expression for the linear predictor is given by: $\eta = \beta_i + \gamma_j + \delta_k$ where indices i, j and k are as follows: $i = 1, 2$ for the two levels of `gender`, $j = 1, 2, 3$ for the three levels of `city`, and $k = 1, 2$ represents the two levels of `blood`.

An alternative, equally acceptable, parametrisation for Model 1 is $\eta = \alpha + \beta_i + \gamma_j + \delta_k$, with indices i, j and k as above, adding the constraint $\beta_1 = \gamma_1 = \delta_1 = 0$, which is the default parametrisation in `R`.

b) [10] According to the default parametrisation in `R` discussed in a), the intercept absorbs the effects of all base levels of the three factors, i.e. represents the value of the linear predictor for a male, with blood type "A", and living in London.

The coefficient denoted by `"blood O"` gives the additional effect of having blood type "0", if compared to the reference group, that is those with blood type "A".

c) [15] The argument that what matters in terms of location is only whether it is in London or not corresponds to supposing that the effects of living in Manchester or Newcastle, compared to London, are the same.

Following again the notation of the default parametrisation of `R` introduced in a), we can express such hypothesis as $H_0 : \gamma_2 = \gamma_3$, which can be tested using the following test statistic:

$$Z = \frac{\widehat{\gamma}_2 - \widehat{\gamma}_3}{se(\widehat{\gamma}_2 - \widehat{\gamma}_3)} \overset{a}{\sim} N(0, 1)$$

The standard error above is given by:

$$\begin{aligned} se(\widehat{\gamma}_2 - \widehat{\gamma}_3) &= \sqrt{[se(\widehat{\gamma}_2)]^2 + [se(\widehat{\gamma}_3)]^2 - 2\,\mathrm{Cov}(\widehat{\gamma}_2, \widehat{\gamma}_3)} \\ &= \sqrt{0.03342^2 + 0.03196^2 - 2 \times 0.000422297} \\ &= \sqrt{0.001293744} = 0.0359687 \end{aligned}$$

so that the observed value of the test statistic is:

$$z_{obs} = \frac{-0.49740 - (-0.34982)}{0.0359687} = -4.1030$$

The hypothesis is rejected if $|z_{obs}| > 1.96$, which is the case, so we cannot assume that the two effects are the same, as argued by the student.

d) [15] The deviance table presents three models, each nested in the following one (top to bottom), starting from the base model, corresponding to the previously presented Model 1, and adding each time a new interaction between two factors.

Model 2 adds to the initial specification of Model 1 the interaction of gender with blood type, while Model 3 adds to Model 2 the interaction of gender with city. A parametric expression for the linear predictors of the proposed models is:

- Model 2: gender * blood + city = gender + blood + city + gender.blood; $\eta = \beta_i + \gamma_j + \delta_k + \zeta_{ik}$ (6 parameters)

- Model 3: gender * (city + blood) = gender + blood + city + gender.blood + gender.city; $\eta = \beta_i + \gamma_j + \delta_k + \zeta_{ik} + \kappa_{ij}$ (8 parameters)

where indices i, j and k are as above.

The number of degrees of freedom for each of the above models is computed as the difference $n-p$, where n is the total number of combinations of factors ($n = 2\times3\times2 = 12$), and p is the number of parameters. The number of parameters is computed starting from Model 1 (5 parameters) and adding for each interaction between two factors an extra $(l-1)(m-1)$ parameters, where l and m represent, respectively, the number of levels of each factor in the interaction.

e) [10] The second row compares Models 2 and 1 as defined earlier, and tests whether the additional interaction gender.blood gives a significant reduction in the deviance. This can be expressed parametrically as $H_0 : \zeta_{ik} = 0$.

Since Model 1 is nested in Model 2, we can test H_0 using the difference between deviances, which follows, in this case, an approximate $\chi^2(1)$ distribution (d.f. correspond to the number of added parameters, as in column 4). The difference in the deviances is displayed in column 5, while the p-value is shown in column 6. Since p-value $= 0.0006258 < \alpha = 0.05$, we reject the null hypothesis, and conclude that the added interaction is sigificant.

The last row compares Models 3 and 2, similarly to what described for row 2. Here $H_0 : \kappa_{ij} = 0$, with p-value $= 0.0000 < \alpha$, so again we reject the null hypothesis, and conclude that the added interaction is sigificant, so that Model 3 is better than Model 2, which was better than Model 1. In conclusion, it is necessary to add interactions to Model 1, and the best alternative is Model 3.

f) [15] Model 4 represents a quasi-likelihood estimation of the expected value of the number of cases of peptic ulcer, where the chosen link function is the canonical link function of the Poisson model, and the variance function is that of the Poisson model too, while the linear predictor is the same as in Model 3. Let Y represent the variable `cases`; we can express the hypotheses as:

- $E(Y) = \mu$
- $\text{Var}(Y) = \varphi\ \mu$
- $\eta = \log(\mu)$

where φ is a dispersion parameter.

The new model is therefore the extension of Model 3 to consider possible over- or underdispersion, as it allows for estimation of the dispersion parameter φ, which was previously set equal to 1. Model 4 gives an estimated value for φ equal to 7.027038, much larger than one.

Not all the estimated parameters are significant, in particular all the interaction coefficients (last three) show a p-value greater than 10%. This shows that not taking into account overdispersion can lead to wrong conclusions in terms of significance of variables.

It is legitimate to use Model 4 for variable `cases`, since such variable is positive; in fact, the hypotheses recalled above imply a positive value for $\mu = \exp(\eta)$ and therefore for the variance, which is a linear function of the mean.

g) [10] The last two rows of the deviance table associated to Model 4 allow us to test the significance of the added interactions `gender:city` and `gender:blood`.

Each test is based on the difference in the scaled deviances between the nested models in two subsequent rows, which follow an approximate F distribution, due to the estimation of parameter φ.

The value of the test statistic and corresponding p-value are shown in the last two columns, showing that in both cases the added interaction is not significant. Therefore this contrasts the evidence found in e), and we conclude that, after accounting for overdispersion, the interactions are no longer significant, so that a model with linear predictor as in Model 1 and overdispersion is the best specification.