Norwegian University of Science and Technology
Department of Mathematical Sciences

Solution TMA4315 GENERALIZED LINEAR MODELS

Friday December 10th, 2010

Problem 1 Number of buss passengers

a) GLM for model 1:

Respnose: $Y_i \sim Po(\lambda_i)$
Assume that the $Y_1, \dots Y_N$ are independent.
$E(Y_i) = \lambda_i = \mu_i$

Link: $\eta_i = \log(\mu_i) \Rightarrow \mu_i = \exp(\eta_i)$

Linear component: $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = X\beta$
where β_0 is the intercept, $x_{i1} = 0$ for non-semester observations and $x_{i1} = 1$ for observations in the semester, and x_{i3} is the temperature for observation i.

Design matrix for $\beta = (\beta_0, \beta_1, \beta_2)^T$:

$$X = \begin{bmatrix} 1 & 1 & 8.8 \\ 1 & 0 & 11.5 \\ 1 & 0 & 12.0 \\ 1 & 1 & 14.8 \\ 1 & 0 & -1.2 \\ 1 & 0 & 7.8 \end{bmatrix}$$

Identifiability: Here corner-stone parametrization is used as we set $x_1 = 0$ for observations with *nonsemester*. An alternative would be to use a sum-to zero constraint

The `R` notation `temp*semester` gives a model with interaction between temperature and semester/non-semester, i.e. the linear component becomes

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} * x_{i2}$$

This means that temperature is allowed to have a different effect (slope) when it is semester (slope $\beta_2 + \beta_3$) or non-semester (slope β_2).

b)
- $\mu = \exp(0.254 + 1.085 - 5.4 \cdot 0.035) = 3.16$
- $\mu = \exp(0.254 + 15.2 \cdot 0.035) = 2.19$

c) Hypothesis:

H_0**:** Model 1 is correct

H_1**:** Model 2 is correct

Likelihood-ratio test: $\Delta D = D_1 - D_2 \sim \chi^2(p_1 - p_2)$ Where D_1 is the deviance for model 1 (with p_1 degrees of freedom) and D_2 is the deviance for model 2 (with p_2 degrees of freedom).
$\Delta D = 17.6777 - 13.981 = 3.696$, and $p_1 - p_2 = 17 - 16 = 1$. And for a test on 5% level we have a critical value of (from table) 3.841, so we keep model 1. (But the test statistic is close to the critical value)

We can decide which of the models *model 1*, *model 2* or *model 3* to use in (at least) two.

1. We have already found that *model 1* is better then *model 2*. We can then do a likelihood-ratio test between *model 1* and *model 3*, and we conclude that we keep *model 1.*
2. We can use AIC, and choose the model with lowest AIC, i.e. model 2.

Both alternatives are correct. The very best answers are the ones that discuss both.

d) Let $Y_1, \dots Y_N$ be independent responses with $Y_i \sim Po(\lambda_i)$, i.e. pdf

$$f(y;\lambda) = \frac{\mu^y}{y!}\exp(-\mu)$$

which gives log-likelihood;

$$l_i(\mu_i) = y_i \ln(\mu_i) - \ln(y_i!) - \mu_i$$

For saturated model (one μ_i per observation y_i); $\delta l_i/\delta\mu_i = 0 \Rightarrow \hat{\mu_i} = y_i$.
For model of interest; fitted value for observation i; $E(Y_i) = \hat{\mu}_i = \hat{y}_i$.
Deviance;

$$\begin{aligned} D &= 2(l_{saturated} - l_{model})) \\ &= 2\sum_{i=1}^{N}(y_i \ln(y_i) - \ln(y_i!) - y_i - (y_i \ln(\hat{y}_i) - \ln(y_i!) - \hat{y}_i)) \\ &= 2\sum_{i=1}^{N}(y_i \ln\frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i)) \end{aligned}$$

Problem 2 Negative binomial distribution

a) If a pdf is a member of the exponential family it can be written as

$$f_y(y;\theta) = \exp(a(y)b(\theta) + c(\theta) + d(y))$$

The probability density function for a negative binomial random variable is

$$f_y(y;\theta,r) = \frac{\Gamma(y+r)}{y!\Gamma(r)}(1-\theta)^r\theta^y$$

and

$$\ln f_y = y\ln(\theta) + r\ln(1-\theta) + \ln(\Gamma(y+r) - \ln y! - \ln\Gamma(r))$$

Hence, f_y belongs to the exponential family with $a(y) = y$, $b(\theta) = \ln(\theta)$, $c(\theta) = r\ln(1-\theta)$ and $d(y) = \ln(\Gamma(y+r)) - \ln y! - \ln\Gamma(r)$. It is of canonical form since $a(y) = y$.

Show that the negative binomial distribution is a member of the exponential family. You can in this question consider r as a known constant.

b) Use the general formulas for a exponential family;

$$E(Y) = \frac{-c'(y)}{b'(\theta)} = r\frac{\theta}{1-\theta}$$

and

$$Var(Y) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{(b'(\theta))^3} = \cdots = \mu\frac{1}{1-\theta}$$

c) GLM for model 1 with negative binomial response:

Respnose: $Y_i \sim nbin(\theta_i, r)$
Assume that the $Y_1, \ldots Y_N$ are independent.
$E(Y_i) = r\frac{\theta}{1-\theta} = \mu_i$
Link: $\eta_i = \log(\mu_i) \Rightarrow \mu_i = \exp(\eta_i)$
Linear component: $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = \beta X$
where βs and x_is as in a).

Need a link-function that ensures positive μ_i, e.g. $\log(\cdot)$.
r is a nuisance parameter, and can be used to fit $Var(Y_i)$ (or adjust $E(Y)$ such that we get the desired $Var(Y_i)$).
With this parametrization of the negative binomial the same sample space. An important difference is that for negative binomial we can have $Var(Y) > E(Y)$. It can therefore be useful if the data are overdisperse.