Norwegian University of Science and Technology
Department of Mathematical Sciences

Solution TMA4315 GENERALIZED LINEAR MODELS

Tuesday December 6th, 2011

Problem 1 Christmas gift preferences

a) GLM for model 1:

Respnose: $Y_i \sim Bin(1, p_i)$
Assume that the $Y_1, \dots Y_N$ are independent.

Logit link: $\eta_i = \log(\frac{p_i}{1-p_i})$

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 females and $x_{i1} = 1$ for males, and x_{i2} is the age for child i.

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

$$X = \begin{bmatrix} 1 & 0 & 7.8 \\ 1 & 1 & 10.5 \\ 1 & 0 & 2.0 \\ 1 & 0 & 8.0 \\ 1 & 1 & 2.0 \\ 1 & 0 & 4.0 \end{bmatrix}$$

Identifiability: Here corner-stone parametrization is used as we set $x_1 = 0$ for $females$. An alternative would be to use a sum-to zero constraint ($beta_{1,male} + \beta_{1,female} = 0$), or (as sex has only two levels and is the only factor) omit the intercept.

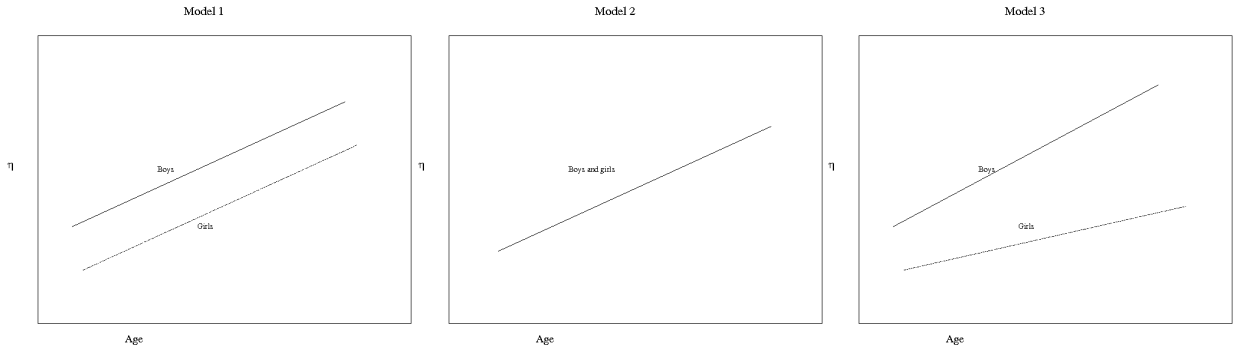
b) In model 1 both sex and age is included. This implies that girls and boys have different models, but the effect of age is the same for both sexes.

In model 2 only age is included, i.e. we have the same model for boys and girls.

In model 3 there are also an interaction between age and sex, hence model 1 in extended such that also the effect of age can differ between the sexes.

The `R` notation `sex*age` gives a model with interaction between sex and age, i.e. the linear component becomes

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



c) For *model 1* is the linear component $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_s$ with estimate $\beta_0 = -0.69613$, $\beta_1 = 0.78254$ and $\beta_2 = 0.06906$. Further is the *logit*-link: $\eta_i = log(\frac{p_i}{1-p_i})$, hence is the probability of preferring hard gifts $p_i = (\frac{\exp(\eta_i)}{1+\exp(\eta_i)})$, and soft gifts $1 - p_i = \frac{1}{1+\exp(\eta_i)}$.

- Probability that a 5 years old ($x_2 = 5$) girl ($x_1 = 0$) prefer soft gifts: 0.59
- Probability that a 15 years old ($x_2 = 15$) girl ($x_1 = 0$) prefer soft gifts: 0.42
- Odds ratio between a 15 years old girl and a 15 year old boy for preferring soft gifts: Odds for 15 years old ($x_{2,girl} = 15$) girl ($x_{1,girl} = 0$) $O_{girl} = \exp(\beta_0 + \beta_1 \cdot 0 + \beta_2 15)$ and for 15 years old ($x_{2,boy} = 15$) boy ($x_{1,boy} = 1$): $O_{boy} = \exp(\beta_0 + \beta_1 \cdot 1 + \beta_2 15)$ $OR = \frac{O_{boy}}{O_{girl}} = \exp(\beta_1) = 2.18$.

d) Models can be compared either using AIC or, if they are nested, likelihood ratio tests. For AIC the lower the better, and between *model1-4*, *model 4* is the best. Likelihood ratio tests are for nested models, for example *model 1* and *model 2*.

Hypothesis:

H_0**:** Model 2 is correct (has fewest parameters)

H_1**:** Model 1 is correct

Likelihood-ratio test: $\Delta D = D_2 - D_1 \sim \chi^2(p_1 - p_2)$ 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).
$\Delta D = 1374 - 1338 = 36$, and $p_1 - p_2 = 998 - 997 = 1$. And for a test on 5% level we have a critical value of (from table) 3.841, so we reject model 2, and conclude that *model 1* fits better.

To evaluate fit for one particular model we can use the sampling distribution of the deviance. For *model 1* the deviance is $D = 1338$ with $\nu = 997$ degrees of freedom. The deviance is approximately $\chi^2(\nu)$-distributed, and $D = 1338$ is in the right tail, i.e. the fit is not very good. (χ^2 with many degrees of freedom is approximately Gaussian with mean ν and standard division $\sqrt{2\nu} = 44.7$.)

To evaluate the models graphically we can plot residuals against fitted values, as well as against covariates. But for binary data this is not very useful, but box plots for residuals for ranges of for example ages might give insight. plot of residuals vs leverage can be useful to find outliers / influential observations.

A next model to fit could either be an extension of *model 3* with and age^2 term to account for non-linearity in the age effect (in *model 4* $agegr2$ has the largest coefficient), or to include an interaction term in *model 4* (as *model 3* is better then *model 1* according to AIC).

e) One possible solution: We can now include electronic gifts as a third response category, and have a multinomial response function $y \sim M(p_H, p_S, p_E)$ where p_H is the probability of preferring hard gifts, p_S soft gifts and p_E electronic gifts. The probabilities have to sum to one; $p_H + p_S + P_E = 1$. We chose to use p_E as reference category, and link functions; $\eta_S = \log(\frac{p_S}{p_H})$ and $\eta_E = \log(\frac{p_E}{p_H})$. Further is each of η_S and η_E given a linear component similar to *model 1*.

Problem 2 Number of Christmas gifts

a) Let $Y_1, \dots Y_N$ be independent responses with $Y_i \sim Po(\mu)$, i.e. probability function

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

A probability function is member of the exponential family if is can be written as;

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

which we can with $a(y) = y$, $b(\mu) = \ln(\mu)$, $c(\mu) = -\mu$ and $d(y) = -\log(y!)$, i.e. it is a member of the exponential family. Since $a(y) = y$ it is also of canonical form.

b) The log likelihood for one observation;

$$l_i(\mu_i) = y_i \ln(\mu_i) - \ln(y_i!) - \mu_i = l_i(\alpha, \beta) y_i(\alpha + \beta x_i) - \ln(y_i!) - \exp(\alpha + \beta x_i)$$

For $n = 20$ observations the likelihood is;

$$L(\alpha, \beta; y_1, \ldots, y_n) = \prod_{i=1}^{n} L_i(\alpha, \beta)$$

and the log-likelihood is

$$l(\alpha, \beta; y_1, \ldots, y_n) = \log(\prod_{i=1}^{n} L_i(\alpha, \beta)) = \sum_{i=1}^{n} (l_i(\alpha, \beta)).$$

The score functions are

$$U(\alpha) = \frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} (-\exp(\alpha + \beta x_i) + y) = \sum_{i=1}^{n} (y_i - \mu_i)$$

and

$$U(\beta) = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} (-\exp(\alpha + \beta x_i) x_i + y) = \sum_{i=1}^{n} (y_i - \mu_i) x_i$$

And hence;

$$U(\alpha, \beta) = \begin{pmatrix} U_1(\alpha, \beta) \\ U_2(\alpha, \beta) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} (Y_i - \mu_i)$$

The information matrix is given by;

$$\mathcal{I} = -E(\begin{pmatrix} \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta} \\ \frac{\partial^2 l}{\partial \alpha \partial \beta} & \frac{\partial^2 l}{\partial \beta^2} \end{pmatrix}) = \sum_{i=1}^{n} \begin{pmatrix} \mu_i & \mu_i x_i \\ \mu_i x_i & \mu_i x_i^2 \end{pmatrix}$$

c) A saturated model (one μ_i per observation y_i, or sometimes defined as one per covariance pattern); $dll_i/d\mu_i = 0 \Rightarrow \hat{\mu}_i = y_i$.
For model of interest; fitted value for observation i; $E(Y_i) = \hat{\mu}_i = \hat{y}_i$.

d) Deviance;

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

There are two parameters in this model (α and β) and 20 data, hence we have $20-2=18$ degrees of freedom.

For Poisson models deviance can be used both for comparing nested models, and (as there are no nuisance parameters) to evaluate fit. See discussion in 1 d).

Problem 3 Valid GLMs Requirements:

Likelihoods: 1. Independent observation

2. Member of the exponential family of canonical form.
3. Should be of same kind for all observations, but possible different expectation/parameter.

Link functions: 1. monotone

2. differentiable

Linear component: 1. linear in parameters

Likelihoods: 1. Gaussian; $Y \sim N(\mu, \sigma^2)$ **OK**

2. Multinomial; $Y \sim M([p_1, p_2, p_3, p_4], N)$ **Not member of exponential family, but OK with Poisson justification. But needs three (4-1) link functions and linear components**
3. Poisson; $Y \sim Po(\theta)$ **OK**

Link functions: 1. $\eta = \cos(\mu)$ **Not monotone**

2. $\eta = \mu$ **OK**
3. $\eta = \log(\mu)$ **OK**

Linear component: 1. $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ **OK**

2. $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$ **OK**
3. $\eta = \beta_0 + \beta_1 x_1 + \beta_1^2 x_2$ **Not linear in** β

The two linear components (1 and 2) can be used for any of the models, which leave us with four alternatives for the link and response:

- Gaussian response with identity link: A common model to used

- Gaussian response with log-link: Gives a model that only has positive expected value (but observations can be negative). Only for special situations.
- Poisson response with identity link. Poisson requires positive expectation, and with this link restrictions has to be imposed on the parameters.
- Poisson response with log link: A common model to use.