TMA4315: Generalised Linear Models Fall 2009
Assignment 4: Models for Multinomial data

Deadline: 24.11.09.

To be delivered in the mailbox of Allessandro Ottavi (Dep. of Mathematical Sciences, 7th floor, SB II) or sent electronically (pdf) to `ottavi@math.ntnu.no`.

Only solutions to Problems 2 and 4 are to be submitted.

Multinomial models

In this seciton we will study models for multinomial data. These can be used when the response variable can be thought of as a realisation of a multinomial distribution with J levels.

The tools to analyse multinomial data are in the **R** library `VGAM` . You need to call this library (`library(VGAM)`) before starting your analysis. The library is already installed on the student server, if you want to install it on your computer you can download it from the web site

`http://www.stat.auckland.ac.nz/~yee/VGAM/`

The function that fits models for multinomial variables is `vglm()` and its syntax is similar to the one of `glm()`. The response variable can be codified as an $n \times J$) matrix of counts where the jth column indicates $Y = j$ and J is the number of values which the response variable Y can assume.

The function `vglm()` returns a series of useful quantities, see `help(vglm)` for details. Suppose you want to extract for example the coefficients of the fitted model `fit`. Then you have two possibilities, call them directly using `fit@coefficients` (Note: in the functions we have used until now the way to call part of the output of a function was `fit$coefficients`!!) or use the **R** function `coef` and type `coef(fit)`.

The two models we will use are the *proportional odds model* and the *multinomial logit model.* These can be fit choosing respectively `family=cumulative(par=T)` and `family=multinomial`.

Exercise 1

The true proportional odds model that explains the dependence of the satisfaction of householders with their present housing circumstances, (Low=1, Medium=2 or High=3, ordered factor,) as a function of the type of rental accommodation (Tower, Atrium, Apartment and Terrace) is the following

$$\begin{aligned}\text{logit}(P(Y \leq 1)) &= 0.1 - [0.1 * I(\text{Apart}) - 0.2 * I(\text{Atr}) + 0.5 * I(\text{Ter}) + 0.4 * I(\text{Tow})] \\ \text{logit}(P(Y \leq 2)) &= 1 - [0.1 * I(\text{Apart}) - 0.2 * I(\text{Atri}) + 0.5 * I(\text{Ter}) + 0.4 * I(\text{Tow})]\end{aligned}$$

where $I()$ is an indicator variable.

a) Simulate a data set from the given model using the **R** function `rmultinom()`. For example you can simulate 40 observation for each value of the covariate and have a data set of 160 observation

b) Fit a proportional odds model to the simulated data using `vglm`. How does the function `vglm` parameterise the factorial explanatory variable? Can you estimate the original parameters in the given model?

c) Fit the same data set using a multinomial model that does not take into account the order of the response variable. Note that in this case the last category of the response variable is taken as the reference. Compare the two fitted models, in particular with respect to the fitted probability values. How large is the loss of not considering the order of the response variable? Try to increase and decrease the number of data in the data set and compare the two models again.

Exercise 2: Models for ordinal responses

The data set `breath.txt` concerns the study of the effect of age and smoking on breathing test results for workers in industrial plants in Texas. The test results have been classified in three categories, namely normal, borderline and abnormal. Thus the response category "breathing result" may be considered an ordinal variable. The registered explanatory variables are age class and whether the subject is a smoker, a former smoker, or has never been smoking.

a) Fit a proportional odds model for the breathing results with only simple effect of age and smoking habits. Is this model satisfactory? Look for example at the residual deviance, the plot of predicted logits against the observed ones, etc.

b) Is an interaction effect between age and smoking habits significant? Check this visually, too.

c) Give an interpretation of the parameters in the fitted model.

Log-linear models

In this section we will study log-linear models. These models can be used when the data under analysis are counts not in form of proportions, a typical example involves counts of events in a Poisson process where the upper limit is unbounded.
The **R** function which we will use to fit log-linear models is `glm()`. To fit a log-linear model you have to choose `family=poisson`, the response variable in this case will be a vector of counts.

Exercise 3

(a) Assume that $Y_i \sim \text{Poisson}(t_i\nu)$ for $i = 1, 2, \ldots, n$.

Show that

$$\hat{\nu} = \frac{\bar{Y}}{\bar{t}}$$

is an unbiased estimator of ν.

Why can we expect that also

$$\hat{\gamma} = \frac{(1/n)\sum_{i=1}^{n}(Y_i - t_i\hat{\nu})^2}{\bar{t}} \approx \nu$$

when n is large?

(*Hint:* Use that $E(Y_i - t_i\nu)^2 = Var(Y_i)$.)

(b) Assume that $Y_i \sim \text{Poisson}(t_i e^{\beta_1+\beta_2 x_i})$ for $i = 1, 2, \ldots, N$.

Suppose x_i can take K different values, $1, 2, \ldots, K$. Then for each k there is a subset $\{Y_i : x_i = k\}$ of variables which satisfy the requirement of (a) above with $\nu = e^{\beta_1+\beta_2 k}$.

This gives us a pair $(\hat{\nu}_k, \hat{\gamma}_k)$ for $k = 1, 2, \ldots, K$. Explain why a plot of these points is supposed to fall on the 45 degrees line through the origin if the model is correct.

(c) Why can we expect a plot of $(k, \ln\hat{\nu}_k)$ for $k = 1, 2, \ldots, K$ to fall near a line too?

(*Hint:* From (b) we would suspect that $\ln\hat{\nu} \approx \ln\nu = \beta_1 + \beta_2 k$.)

Exercise 4

Consider the data set `Insurance` contained in the **R** library `MASS`. To get more information of the data set type `help(Insurance)`. The response variable, numbers of claims Y_i, can be considered Poisson distributed $Y_i \sim Poisson(\mu_i)$.

The aim is to study the dependence between the number of claims and the other registered variables.

The variable `Holders` (the number of people who hold a policy) is a measure of "risk exposure" of the insurance company (the larger the number of people who hold a policy the larger the number of claims). Therefore it cannot be treated as an ordinary covariate, but it has to be taken into account in order to get a reasonable model. We model the response variable as

$$Y_i \sim Poisson(t_i \exp \mathbf{x}_i^T \beta)$$

where t_i is the number of policy holders, $\mathbf{x}_i$ a vector of explanatory variables and β a vector of unknown parameters to be estimated. Note that we do not wish to estimate a parameter for t_i, and we therefore need to specify this variable as an **offset** in the following way:

```
model=glm(Claims ~ District + Group + Age, offset=log(Holders),
family=poisson, data=Insurance)
```

(a) Fit a model where the explanatory variables are treated as factors, and where no interactions are present.

Then fit the model where there is interaction between the factors Group and Age. Use deviances computed by R to compare the two models. Which one will you prefer?

(b) For the preferred model, compute Pearson's X^2 and compare with the value of the deviance. Is the model fit satisfactory?

(c) The explanatory variables above have been treated as factors with nominal levels. However, both Group and Age are defined in terms of meaningful quantities and could therefore possibly be represented as numerical covariates. As a start one may therefore try to replace each of Group and Age by variables with values 1,2,3,4.

Write down the resulting full GLM-model with no interactions, when District still is taken as a factor while Group and Age are taken as numerical covariates. What will be the benefit of such a model compared to the ones considered above?

(d) To make a rough check whether the variables Group and Age can be entered in the model with only linear terms of the form, say, "$\beta_G Group$" and "$\beta_A Age$" we can use the plot suggested in Exercise 3 (c), considering Group and Age separately in models with single explanatory variables. Explain why this is so and make the plots using R! What is your conclusion?