Norwegian University of Science and Technology
Department of Mathematical Sciences

Solution TMA4315 GENERALIZED LINEAR MODELS

Tuesday December 13th, 2012

Problem 1 Precipitation in Trondheim tomorrow?

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: Model 1: $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_3 = X_1\beta$ where x_{i1} is the amount in forecast for day i, $x_{i2} = 1$ if precipitation in forecast, and is zero otherwise. β_3 is a vector $\beta_3 = \{\beta_{3,OF=0}, \beta_{3,OF=1}, \beta_{3,OF=2}, \beta_{3,OF=3}\}$ and x_{i3} is a vector of length four of zeros except for the element corresponding to OF_1 which is one. Model with $Fore$ as covariate, and $ForeBin$ and OF as factors.

Model 2: $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_3 = X_2\beta$ where β_0 is an intercept, x_{i1} and $x_{i2} = 1$ as for Model 1, and β_3 and x_{3i} are now a scalars. Model with $Fore$ and OF as covariates, and $ForeBin$ as factor.

Model 3: $\eta_i = \beta_0 + \beta_1 x_{i1} = X_3\beta$ where β_0 is an intercept, x_{i1} as for Model 1. Model with $Fore$ as covariates.

Discussion assumptions: Independence is the most critical assumptions. The weather today and tomorrow are not independent, but as the model here includes the forecast (in different versions), it is closer to being the error in the forecast that should be independent. Also note that OF is really an interaction term between yesterday's precipitation forecast and occurrence. Hence, model 2 includes yesterdays weather, so much of the temporal should be included from there.

Model 1 include OF as a factor (with 4 levels), while Model 2 include OF as a covariate.

Identifiability: Here corner-stone parametrization is used as we set $x_1 = 0$ for no precipitation for model 1 and model 2. For model 1 OP is treated as a factor, and identifiability is ensured by omitting the intercept.

Design matrix for model 1:

$$X_1 = \begin{bmatrix} 3.0 & 1 & 0 & 0 & 0 & 1 \\ 0.5 & 1 & 0 & 0 & 0 & 1 \\ 0.0 & 0 & 0 & 0 & 1 & 0 \\ 0.0 & 0 & 1 & 0 & 0 & 0 \\ 0.0 & 0 & 1 & 0 & 0 & 0 \\ 0.0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$X_2 = \begin{bmatrix} 1 & 3.0 & 1 & 3 \\ 1 & 0.5 & 1 & 3 \\ 1 & 0.0 & 0 & 2 \\ 1 & 0.0 & 0 & 0 \\ 1 & 0.0 & 0 & 0 \\ 1 & 0.0 & 0 & 0 \end{bmatrix}$$

$$X_3 = \begin{bmatrix} 1 & 3.0 \\ 1 & 0.5 \\ 1 & 0.0 \\ 1 & 0.0 \\ 1 & 0.0 \\ 1 & 0.0 \end{bmatrix}$$

b) According to *model 1*: What is the probability for precipitation if it the forecast is 5mm and $OF = 0$?
$\eta_i = 1.56$, $p_i = (\frac{\exp(\eta_i)}{1+\exp(\eta_i)}) = 0.83$

According to *model 2*: What is the probability for precipitation if it the forecast is 5mm and $OF = 3$?
$\eta_i = 1.29$, $p_i = (\frac{\exp(\eta_i)}{1+\exp(\eta_i)}) = 0.78$

According to *model 3*:What is the odds ratio between a day with forecast 0mm and a day with forecast 5mm?
$O_{Fore=5} = \exp(\beta_0 + \beta_1 \cdot 5)$ $OR = \frac{O_{Fore=5}}{O_{Fore=0}} = \exp(\beta_1(5-0)) = 28.2.$

c) Model 1 and model 3 are nested, and we can use a likelihood ratio test.

Likelihood ratio tests are for nested models, for example *model 1* and *model 2*.

Hypothesis:

H_0**:** Model 3 is correct (has fewest parameters)

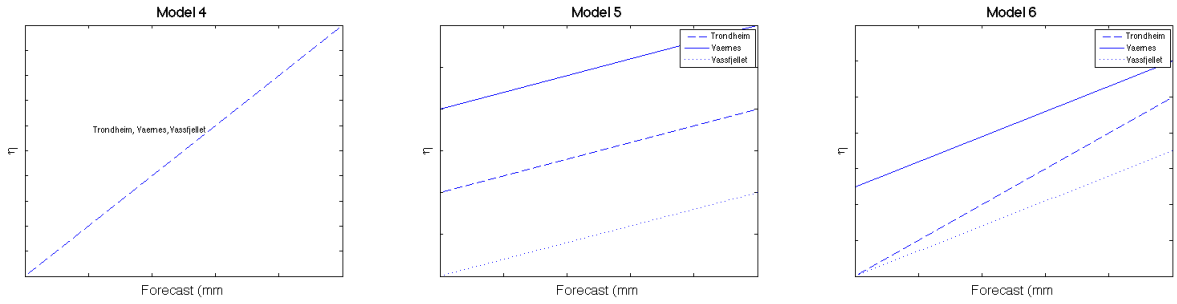
H_1**:** Model 1 is correct

Likelihood-ratio test: $\Delta D = D_3 - D_1 \sim \chi^2(p_1 - p_3)$ Where D_1 is the deviance for model 1 (with p_1 parameters) and D_3 is the deviance for model 3 (with p_3 parameters).
$\Delta D = 101.09 - 104.91 = 3.7$, and $p_1 - p_3 = 6 - 2 = 4$. And for a test on 5% level we have a critical value of (from table) 9.5, so we can not reject model 3, and conclude that *model 3* fits better.

Model 2 and model 3 are also nested, and a log-ratio test of these also result in not rejecting H_0 : Model 3 is correct.

AIC can also be used for comparing models. For AIC the lower the better, and between *model1-3*, also AIC indicates that *model 3* is the best.

d) Model 4: Same model for all locations.
Model 5: Same linear relation of forecast for all models.
Model 6: Each location has their own model.



One assumption of GLMs is that the observations are independent. As we now have data for three close locations, it is reasonable that they are dependent. For example if the forecast gave no precipitation, but a weather-system containing precipitation hits, it will often give precipitation at all locations. Hence, the assumption of independence should at least be tested.

An alternative model would be a GLMM with day number as a random effect.

Problem 2 Precipitation in Trondheim as snow, sleet or rain?

a) As the kind of precipitation goes from snow, to sleet to rain with increasing temperature, it is reasonable to use a model for ordinal data, e.g. a proportional odds ordinal odds model.

The model should be set up, linear components written out and parameters interpreted.

Problem 3

a) Member of exponential family if

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

Set in for $\alpha = \alpha_i$ and $\beta = \mu_i/\alpha_i$, and get

$$f_Y(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta) = \exp(a(y)b(\mu_i) + c(\mu_i) + d(y))$$

with
$a(y_i) = y_i$
$b(\mu_i) = -\alpha_i/\mu_i$
$c(\mu_i) = -\ln((\mu_i/\alpha_i)^{\alpha_i}\Gamma(\alpha_i)) = -\alpha_i \ln(\mu_i) + \alpha_i \ln(\alpha_i) - \ln(\Gamma(\alpha_i))$
$d(y_i) = (\alpha_i - 1)\ln(y_i)$

I.e. member of the exponential family of canonical for as $a(y) = y$.

This gives us $E(Y_i) = b'(\mu_i)/c'(\mu_i) = \cdots = \mu_i$ and $Var(Y_i) = \cdots = \mu_i^2/\alpha_i$.

Interpretation of α_i: We see that the standard deviation is proportional to the expected value, and that α is a scaling parameter for the variance.

b) A saturated model is a model with as many paramters as possible. In most cases one parameter per observation. Log-likelihood for one observation:

$$l_i = \log(f_Y(y)) = y_i b(\mu_i) + c(\mu_i) + d(y_i)$$

To find maximum likelihood for the saturated model we take the derirative with respect to μ_i, set equal to zero, and solve for μ_i.

$$y_i b'(\mu_i) + c'(\mu_i) = 0 \cdots \Rightarrow \hat{\mu}_{imax} = y_i$$

Let $\hat{\mu}_i$ be the fitted value from our model. The deviance is;

$$\begin{aligned} D &= 2(l_{saturated} - l_{model})) \\ &= \ldots \\ &= -2\sum_{i=1}^{N} \alpha_i \ln(y_i/\hat{\mu}_i) + \alpha_i \frac{\hat{\mu}_i - y_i}{\hat{\mu}_i} \end{aligned}$$

c) It is reasonable to use gamma distribution for amount of precipitation given precipitation. From a) and b) we have learned that it is member of the exponential family, and hence it can be used as response function for a GLM. Gamma distributions require a positive expected value, and hence a log-link (or an other monotone differential function that ensures positive μ) should be used. For the linear component $\eta_i = \beta_0 + \beta_1 x_i$ is one alternative. But as a log link is used $E(Y_i) = exp(\eta_i)$, and $\eta_i = \beta_0 + \beta_1 \log(x_i)$ is an other natural choice.