Introduction on to Generalized Linear Models (GLM)

STK3100/STK4100 - August 22th 2016

Sven Ove Samuelsen/Anders Rygh Swensen

Department of Mathematics, University of Oslo

2015
Plan for first lecture:

1. Introduction, Literature, Program
2. Examples
3. Informal definition of GLM
4. Some extensions of GLM
5. Plan for the course
Introduction

- The topic of generalized linear models (with extensions) is central classes of more complicated, but standard models beyond multiple regression / anova.
- In particular we will see how binary data, data on counts, categorical (multinomial) data and longitudinal/panel data can be analyzed in a regression (like) setting.
- The purpose of the course is twofold: first to see how these models can be in real applications but also to understand the mathematical background for the models.
Textbook (literature)


Web page: http://www.actuary.mq.edu.au/research/books/GLMsforInsuranceData

Many data sets we will use can be found here.

As earlier we will use data set from many settings: medicine / biology, social science/ economics/ engineering . But a large part will come from insurance.
Textbook (literature), cont.

Textbook for the Generalized Linear Mixed Models, GLMM:
Textbook (literature), cont.

The book is available in the science library.
In the course the R package downloadable from http://mirrors.sunsite.dk/cran/ will be used. It runs under the most common operative systems. Most of the time procedures and functions available in R will be used. Not much own programming will be necessary. For a short introduction to R, see the web page of STK1110 last year, STK1110-h15. A fine overview of R is the book written by Peter Dahlgaard: Introductory Statistics with R, 2nd ed., 2008, Springer.
Example 1: Birth weight and length of pregnancy

<table>
<thead>
<tr>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (weeks)</td>
<td>Birth weight (gram)</td>
</tr>
<tr>
<td>40</td>
<td>2968</td>
</tr>
<tr>
<td>38</td>
<td>2795</td>
</tr>
<tr>
<td>40</td>
<td>3163</td>
</tr>
<tr>
<td>35</td>
<td>2925</td>
</tr>
<tr>
<td>36</td>
<td>2625</td>
</tr>
<tr>
<td>37</td>
<td>2847</td>
</tr>
<tr>
<td>41</td>
<td>3292</td>
</tr>
<tr>
<td>40</td>
<td>3473</td>
</tr>
<tr>
<td>37</td>
<td>2628</td>
</tr>
<tr>
<td>38</td>
<td>3176</td>
</tr>
<tr>
<td>40</td>
<td>3421</td>
</tr>
<tr>
<td>38</td>
<td>2975</td>
</tr>
</tbody>
</table>

Average: 38.33 3024.00 38.75 2911.33

Of interest is the growth per week at the end of the pregnancy and if there is any difference between boys and girls.
Introduction on to Generalized Linear Models (GLM)

Scatter plot for Ex 1.

Length of pregnancy (weeks)

Weight (g)

● boys

● girls
Example 2: Lethal dose for beetles

Around 480 beetles were exposed for eight different concentrations of \( \text{CS}_2 \). The number of deaths for the various concentrations were recorded.

<table>
<thead>
<tr>
<th>Dose ((\log_{10} \text{CS}_2 \text{mg l}^{-1}))</th>
<th>No</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6907</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>1.7242</td>
<td>60</td>
<td>13</td>
</tr>
<tr>
<td>1.7552</td>
<td>62</td>
<td>18</td>
</tr>
<tr>
<td>1.7842</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>1.8113</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>1.8369</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>1.8610</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>1.8839</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

What is the relation of size of dose and mortality?
Proportion dead beetles in Ex 2.
Example 3: number of children among pregnant.

de Jong & Heller, page 15-16: Data for number of children among 141 pregnant women of different ages.
The number increases with age, see figure 1.11 i deJ&H
Example 3b: number of third party claims

de Jong & Heller, side 17: Data over number of claims in 176 geographical regions in New South Wales in a 12-months period.

Explanative variables, covariates:

- Statistical category, 13 categories
- Number of accidents in the region
- Number of killed and injured
- Size of population

In both examples: the response may be Poisson distributed.
Typical model for Ex 1: Linear regression

For $k = 1, \ldots, 12$ and $j = 1, 2$ (where $j = 1$ denotes boy and $j = 2$ denotes girl)

\[
y_{jk} = \text{birth weight for baby nr. } k \text{ gender nr. } j
\]
\[
x_{jk} = \text{length of pregnancy for baby nr. } k \text{ gender nr. } j
\]

assume

\[
y_{jk} = \alpha_j + \beta x_{jk} + \epsilon_{jk}
\]

where $\epsilon_{jk} \sim N(0, \sigma^2)$, i.e. normally distributed with expectation 0 and same variance $\sigma^2$ and also independent.
Parameters in the linear part:

\[ \beta = \text{slope} \]
\[ \alpha_j = \text{intercept for gender } j \]
Least squares fit for Example 1.

Estimates: $\hat{\alpha}_1 = -1447$, $\hat{\alpha}_2 = -1610$, $\hat{\beta} = 121$
Alternative formulation Ex. 1

- Linearity: $E[y_{jk}] = \mu_{jk} = \alpha_j + \beta x_{jk}$
- Constant variance: $\text{Var}[y_{jk}] = \sigma^2$
- Normality assumption: $y_{jk} \sim N(\mu_{jk}, \sigma^2)$
- Independent responses: $y_{jk}$’s independent
Alternative formulation Ex. 1, cont

I GLM (and STK3100) three first features are modified to

- Linearity after transformation via "link-function" $g()$:
  \[ g(\mu_{jk}) = \alpha_j + \beta x_{jk} \]

- Variance may depend on the expectation of the responses.

- Other distributions for the responses: Binomial, Poisson, Gamma, ...

But independent responses are still assumed.
EX. 2: Mortality of beetles

It is reasonable to assume \( y_i = \) number dead beetles for dose \( x_i \) is binomially distributed. \( y_i \sim \text{bin}(n_i, \pi_i) \)

where \( \pi_i = \) probability for beetle dying with dose \( x_i \) and \( n_i = \) number of beetles receiving dose \( x_i \).

Linear model for \( \pi_i \) fitted with least squares problematic because

- \( 0 \leq \pi_i \leq 1 \) in contrast to expression \( \alpha + \beta x_i \)
- \( \text{Var}(y_i) = n_i \pi_i (1 - \pi_i) \), i.e. non-constant (heteroskedastic) structure of variance
Usual model for Ex. 2: Logistic regression

Logistic model of regression:

\[ \pi_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \]

Then \( 0 \leq \pi_i \leq 1 \)

Fit the logistic model of regression with Maximum Likelihood (ML).

- Takes into account binomially distributed responses (and non-constant variance)
- Efficient estimators (approximately for large data)
Logistic regression for Ex. 2: Number of dead beetles

MLE: $\hat{\alpha} = -60.72, \hat{\beta} = 34.27$

Predicted probabilities: $\hat{\pi} = \frac{\exp(\hat{\alpha} + \hat{\beta} x)}{1 + \exp(\hat{\alpha} + \hat{\beta} x)}$
Estimating logistic regression

Storvik: "Numerical optimization of likelihoods: Additional literature for STK2120" describes a Newton-Raphson routine in \( R \) for fitting logistic regression to such observations. This is already implemented in \( R \). Use commando

\[
\text{glm(cbind(Dead, No-Dead)~Dose, family=binomial)}
\]
Example of GLM

- \texttt{glm} = Generalized Linear Model
- \texttt{family=}binomial because data binary or binomial.
- For binomial data \texttt{cbind(Dead, No-Dead)} needs "no. successes" (dead) and "no. failures" (No-Dead).
Ex. 3: number of previous children for mothers

\[ y_i = \text{number of previous children for mother } i. \]
Reasonable to assume \( y_i \) Poisson distributed with expectation \( \mu_i \)
where \( \mu_i \) depends on \( x_i = \text{mothers age}. \)

As in Ex. 2:

- Expectations \( \mu_i > 0 \)
- Variance of \( y_i \) equals \( \mu_i \), i.e. non-constant variance
Usual solution: Poisson-regression

\[ y_i \sim \text{Po}(\mu_i) \text{ where } \mu_i = \exp(\alpha + \beta x_i) \]

This is also a GLM and can be fitted with the \texttt{glm}-routine.

Only have to specify that data is assumed to be Poisson distributed with \texttt{family=poisson}
MLE for $(\alpha, \beta)$: $(\hat{\alpha}, \hat{\beta}) = (-4.0895, 0.1129)$

Fitted expectations: $\hat{\mu}_i = \exp(\hat{\alpha} + \hat{\beta}x_i)$
Definition of GLM

Independent responses $y_1, y_2, \ldots, y_n$

Vectors of covariates $x_1, x_2, \ldots, x_n$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is $p$-dimensional

A GLM = Generalized Linear Model is defined by the following three components:
Definition of GLM, cont.

- $y_1, y_2, \ldots, y_n$ has a distribution belonging to an exponential family
  (Exponential families will be defined later, suffices to know that normal-, binomial-, Poisson-, gamma-distributions belong to the exponential family)
- Linear components (predictors) $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- Link function $g()$: The expectation $\mu_i = \text{E}[y_i]$ is related to the linear component by $g(\mu_i) = \eta_i$
Linear regression is a GLM

- Responses ($y_i$-er) from normal distribution
- Linear component $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- $E[y_i] = \mu_i = \eta_i$, i.e link function $g(\mu_i) = \mu_i$ is the identity function

In particular R-commands `lm` for linear regression and `glm` essentially the same, only a bit different output.

Linear regression is in particular default-specification of `glm`
Ex. 1: Birth weights

```r
> lm(vekt~sex+svlengde)

Call:
  lm(formula = vekt ~ sex + svlengde)

Coefficients:
(Intercept)         sex        svlengde
   -1447.2      -163.0        120.9
```
Ex. 1: Birth weights

```r
> glm(vekt~sex+svlengde)

Call: glm(formula = vekt ~ sex + svlengde)

Coefficients:
(Intercept) sex svlengde
   -1447.2  -163.0   120.9

Degrees of Freedom: 23 Total (i.e. Null); 21 Residual
Null Deviance: 1830000
Residual Deviance: 658800 AIC: 321.4
```

Sven Ove Samuelsen/Anders Rygh Swensen

Introduction on to Generalized Linear Models (GLM)
Logistic regression is a GLM

- Responses ($y_i$'s) binomially distributed bin($n_i, \pi_i$)
- Linear component $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- $E[y_i]/n_i = \pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$, so that link function $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

Denote $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi)$ as logit-function.
Logistic regression is GLM, cont.

```r
> glm(cbind(Dode,Ant-Dode)~Dose,family=binomial)

Call:  glm(formula = cbind(Dode, Ant - Dode) ~ Dose, family = binomial)

Coefficients:
(Intercept)       Dose
       -60.72       34.27

Degres of Freedom:  7 Total (i.e. Null);  6 Residual
Null Deviance:      284.2
Residual Deviance:  11.23          AIC: 41.43
```
Poisson regression is a GLM

- Response \( y_i \sim \text{Po}(\mu_i) \)
- Linear component \( \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \)
- \( \text{E}[y_i] = \mu_i = \exp(\eta_i) \), i.e. link function \( g(\mu_i) = \log(\mu_i) \) is the (natural) logarithmic function.
Introduction on to Generalized Linear Models (GLM)

Poisson regression is a GLM, cont.

> glm(children~age, family=poisson)

Call: glm(formula = children ~ age, family = poisson)

Coefficients:
(Intercept)   age
     -4.0895    0.1129

Degrees of Freedom: 140 Total (i.e. Null); 139 Residual
Null Deviance: 194.4
Residual Deviance: 165   AIC: 290
Other GLM’s:

- Count data with negative binomial distribution: Over dispersion.
- Continuous, non-normal responses: gamma-, inverse gaussian distributions

These will be considered.
Some other extensions

Extensions of GLM:

- Multinomial responses (ordinal and nominal)
- Life time data
- Analysis of dependent data, GLMM
- Generalized additive models (GAM)

We will consider multinomial responses and GLMM.
Example 4: Growth of trees and ozone exposure

Growth for two groups of trees is recorded at five different points of time. Of the trees 54 are located in an environment with heavy traffic and 25 trees are a control group. In total there are 395 measurements $y_{i,j}$, $i = 1, \cdots, 79$, $j = 1, \cdots, 5$. 
Plot of 10 profiles in each group
Introduction on to Generalized Linear Models (GLM)
Exposure of ozone

Linear mixed-effects model fit by maximum likelihood

Random effects:
Formula: ~1 + Time | tree
Structure: General positive-definite, Log-Cholesky parametrization

<table>
<thead>
<tr>
<th></th>
<th>StdDev</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.7909</td>
<td>(Intr)</td>
</tr>
<tr>
<td>Time</td>
<td>0.0025</td>
<td>-0.649</td>
</tr>
<tr>
<td>Residual</td>
<td>0.1626</td>
<td></td>
</tr>
</tbody>
</table>

Fixed effects: size ~ Time + factor(treat) * Time

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.1217</td>
<td>0.1780</td>
<td>314</td>
<td>11.91</td>
<td>0.000</td>
</tr>
<tr>
<td>Time</td>
<td>0.0141</td>
<td>0.0006</td>
<td>314</td>
<td>22.32</td>
<td>0.000</td>
</tr>
<tr>
<td>factor(treat)</td>
<td>0.2217</td>
<td>0.2154</td>
<td>77</td>
<td>1.03</td>
<td>0.307</td>
</tr>
<tr>
<td>ozone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time:factor(treat) ozone</td>
<td>-0.0021</td>
<td>0.0008</td>
<td>314</td>
<td>-2.79</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Survey of textbook by de Jong & Heller

- **Chapter 1**: Introduction, Data examples: will not be treated in detail
- **Chapter 2**: Diverse distributions: with some exceptions known before
- **Chapter 3**: Exponential classes, ML-estimation
- **Chapter 4**: Linear modeling (mainly known from STK1110/STK2120)
- **Chapter 5**: Generalized linear models
Survey of textbook by de Jong & Heller, cont.

- Chapter 6: Count data (Poisson regression, overdispersion)
- Chapter 7: Categorical responses (binomial data, multinomial data)
- Chapter 8: Continuous responses
- Chapter 9: Correlated data
- Chapter 10: Extensions
Plan for course, STK3100/STK4100

Will follow the textbook of de Jong & Heller, but not in all details, and not in sequence. Also some parts must be supplemented. In the last part of the course we will treat GLMM and the relevant material in Zuur et al.

Approximate plan for first lectures:
Introduction, today!

Chapter 4. Linear models, mainly repetition of STK1110/STK2120, Friday August 26th

Chapter 3: Exponential classes, August 29th.

Chapter 5: GLM and ML-theory September 5th.

Chapter 7: Binomial and binary data

Chapter 6: Count data
Plan for course, STK3100/STK4100, cont.

- Chapter 7: Multinomial data
- Chapter 8: A little of continuous responses
- Extensions: Correlated data and GAM, material from Zuur et al. chapters 6, 7 and 13.