TMA4315 Generalized linear models H2018

Module 9: SUMMING UP

Mette Langaas, Department of Mathematical Sciences, NTNU

22.11.2018 [PL]

Contents

(Latest changes: 22.11 second version).

Overview

Topics

- course content and learning outcome
- reading list
- course topic and modules
    - core concepts: exponential family, models: LM/GLM/mvGLM/LMM/GLMM, likelihood, maximum likelihood, score vector, Fisher information, Fisher scoring, Wald/LRT(/score) tests, deviance, AIC

    - incoming questions: overview of models (including categorical regression), exponential family and canonical link (why?), likelihood-score-Fisher information, which tests for what, model assessment (deviance and residuals)
- exam and exam preparation
- suggestions for statistics-related courses in year 4 and 5
- questionnaire

Classnotes:

- Notes from overview
- Overview sheet (made before class)
- Basic 5 (made before class)
- Notes from the print-out for interpreting LM/GLM/VGLM/LMM/GLMM (made before class)
- Notes from "problem types on the written exam" (made before class)

About the course

Content

Univariate exponential family. Multiple linear regression. Logistic regression. Poisson regression. General formulation for generalised linear models with canonical link. Likelihood-based inference with score function and expected Fisher information. Deviance. AIC. Wald and likelihood-ratio test. Linear mixed effects models with random components of general structure. Random intercept and random slope. Generalised linear mixed effects models. Strong emphasis on programming in R. Possible extensions: quasi-likelihood, over-dispersion, models for multinomial data, analysis of contingency tables, quantile regression.

H2018: Lectured multinomial data (categorical regression), but did still not cover so much over-dispersion, and *did not cover* quasi-likelihood, contingency tables and quantile regression (of the possible extensions).

Learning outcome

Knowledge

The student can assess whether a generalised linear model can be used in a given situation and can further carry out and evaluate such a statistical analysis. The student has substantial theoretical knowledge of generalised linear models and associated inference and evaluation methods. This includes regression models for normal data, logistic regression for binary data and Poisson regression. The student has theoretical knowledge about linear mixed models and generalised linear mixed effects models, both concerning model assumptions, inference and evaluation of the models. Main emphasis is on normal, binomial and Poisson models with random intercept and random slope.

Skills

The student can assess whether a generalised linear model or a generalised linear mixed model can be used in a given situation, and can further carry out and evaluate such a statistical analysis.

Final reading list

Fahrmeir, Kneib, Lang and Marx (2013): Regression, Springer: eBook (free for NTNU students). https://link.springer.com/book/10.1007%2F978-3-642-34333-9

- Chapter 2: 2.1, 2.2, 2.3, 2.4, 2.10
- Chapter 3 (also on reading list for TMA4267)
- Chapter 5: 5.1, 5.2, 5.3, 5.4, 5.8.2
- Chapter 6: but not p 344-345 nominal models and latent utility models, not 6.3.2 Sequential model, and not category specific varables on page 344-345.
- Chapter 7: 7.1, 7.2, 7.3, 7.5, 7.7, 7.8.2. [In greater detail: pages 349-354 (not "Alternative view on the random intercept model"), 356-365 (not 7.1.5 "Stochastic Covariates""), 368-377 (not "Bayesian Covariance Matrix""), 379-380 (not "Testing Random Effects or Variance Parameters"", only last part on page 383), 383 (middle), 389-394, 401-409. Note: Bayesian solutions not on the reading list.]
- Appendix B.1, B.2, B.3 (not B.3.4 and B.3.5), B.4

In addition to the Fahrmeir et al book, on the reading list is also:

- All the 9 module pages (but module 1 and 9 does not have theory that is not in 2-8).
- The three compulsory exercises.

The modules

1. Introduction (exponential family, Rstudio, ggplot and R Markdown)
2. Multiple linear regression (emphasis on likelihood)
3. Binary regression (independent responses, binary individual and grouped response)
4. Count and continuous positive reponse data (independent responses, Poisson- and gamma regression)
5. Generalized linear models: common core
6. Categorical regression (multinomial distribution, multivariate GLM, nominal and ordinal models)
7. Linear mixed models (normal response, clustered data or repeated measurements)
8. Generalized mixed effects models (non-normal response, clustered data or repeated measurements)
9. Summing-up (this module)

Core of the course: regression

Main question: what it the effect of covariate(s) x on the response(s) y?

Examples

- [M2] Munich rent index
- [M3] Mortality of beetles, infant respiratory disease, contraceptive use.
- [M4] Female crabs with satellites, smoking and lung cancer, time to blood coagulation, precipitation in Trondheim, treatment of breast cancer.
- [M6] Alligator food, mental health.
- [M7+8] Richness of species at beaches, sleep deprivation, trawl fishing.

The five ingredients

1. **Model specification**: an equation linking (conditional) mean of the response to and the explanatory variables, and a probability distribution for the response. We only consider responses from exponential family.
   a. multiple linear regression model (normal response)
   b. univariate generalized linear model (normal, binomial, Poisson, gamma)
   c. multivariate generalied linear model (multinomical: nominal and ordinal)
   d. linear mixed effect models (normal response, correlated within clusters)
   e. generalized linear mixed models (binomial, Poisson)

---

2. **Likelihood** - used to estimate parameters (ML and a bit on REML): score function, Fisher information, Fisher scoring (IRWLS).
3. **Asymptotic distribution** of maximum likelihood estimators (multivariate normal) and tests (chisquared).
4. **Inference**: interpretation of results, plotting results, confidence intervals, hypothesis tests (Wald,LRT,score).
5. Checking the **adequacy of the model** (deviance, also residuals, qqplots - but very little focus in our course *outside the normal model*), **choose between models** (nested=LRT or AIC, not nested=AIC),

Understanding and comparing R print-outs

The print-outs below are from LM `lm`, GLM `glm`, mvGLM `vglm`, LMM `lmer` and GLMM `glmer`.

For H2018: at the exam venue the `vglm`, `lmer` and `glmer` functions are not available (not installed), this means that you do not need to be able to run analyses with these, but you still need to be able to interpret output!

Below we have fit a model to a data set, and then printed the `summary` of the model. For each of the print-outs you need to know (be able to identify and explain) every entry. In particular identify and explain:

- which model: model requirements
- how is the model fitted (versions of maximum likelihood)
- parameter estimates for β
- inference about the β: how to find CI and test hypotheses (which hypothesis is reported test statistic, and possibly p-value for)
- model fit (deviance, AIC, R-squared, F)

In addition, further inference can be made using `anova(fit1,fit2)`, `confint`, `residuals`, `fitted`, `AIC`, `ranef` and other functions we have worked with in the PL, IL and on the compulsory exercises.

MLR - multiple linear regression

```r
library(gamlss.data)
fitLM = lm(rent ~ area + location + bath + kitchen + cheating, data = rent99)
summary(fitLM)
fitGLM = glm(rent ~ area + location + bath + kitchen + cheating, data = rent99)
summary(fitGLM)
```

```
##
## Call:
## lm(formula = rent ~ area + location + bath + kitchen + cheating,
##     data = rent99)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -633.41  -89.17   -6.26   82.96 1000.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.9733    11.6549  -1.885   0.0595 .
## area          4.5788     0.1143  40.055  < 2e-16 ***
## location2    39.2602     5.4471   7.208 7.14e-13 ***
## location3   126.0575    16.8747   7.470 1.04e-13 ***
## bath1        74.0538    11.2087   6.607 4.61e-11 ***
## kitchen1    120.4349    13.0192   9.251  < 2e-16 ***
## cheating1   161.4138     8.6632  18.632  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 145.2 on 3075 degrees of freedom
## Multiple R-squared:  0.4504, Adjusted R-squared:  0.4494
## F-statistic:   420 on 6 and 3075 DF,  p-value: < 2.2e-16
##
##
## Call:
## glm(formula = rent ~ area + location + bath + kitchen + cheating,
##     data = rent99)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -633.41   -89.17    -6.26    82.96  1000.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.9733    11.6549  -1.885   0.0595 .
## area          4.5788     0.1143  40.055  < 2e-16 ***
## location2    39.2602     5.4471   7.208 7.14e-13 ***
## location3   126.0575    16.8747   7.470 1.04e-13 ***
## bath1        74.0538    11.2087   6.607 4.61e-11 ***
## kitchen1    120.4349    13.0192   9.251  < 2e-16 ***
## cheating1   161.4138     8.6632  18.632  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 21079.53)
##
```

```
##     Null deviance: 117945363  on 3081  degrees of freedom
## Residual deviance:  64819547  on 3075  degrees of freedom
## AIC: 39440
##
## Number of Fisher Scoring iterations: 2
```

GLM - Binomial regression with logit-link

```r
library(investr)
fitgrouped = glm(cbind(y, n - y) ~ ldose, family = "binomial", data = investr::beetle)
summary(fitgrouped)
```

```
##
## Call:
## glm(formula = cbind(y, n - y) ~ ldose, family = "binomial", data = investr::beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## ldose         34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

GLM - Poisson regression with log-link

```r
crab = read.table("https://www.math.ntnu.no/emner/TMA4315/2018h/crab.txt")
colnames(crab) = c("Obs", "C", "S", "W", "Wt", "Sa")
crab = crab[, -1]  #remove column with Obs
crab$C = as.factor(crab$C)
model3 = glm(Sa ~ W + C, family = poisson(link = log), data = crab, contrasts = list(C = "contr.sum"))
summary(model3)
```

```
##
## Call:
## glm(formula = Sa ~ W + C, family = poisson(link = log), data = crab,
##     contrasts = list(C = "contr.sum"))
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0415  -1.9581  -0.5575   0.9830   4.7523
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.92089    0.56010  -5.215 1.84e-07 ***
## W            0.14934    0.02084   7.166 7.73e-13 ***
## C1           0.27085    0.11784   2.298   0.0215 *
## C2           0.07117    0.07296   0.975   0.3294
## C3          -0.16551    0.09316  -1.777   0.0756 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 559.34  on 168  degrees of freedom
## AIC: 924.64
##
## Number of Fisher Scoring iterations: 6
```

Categorical regression, nominal model

```r
# data from Agresti (2015), section 6, with use of the VGAM packages
data = "http://www.stat.ufl.edu/~aa/glm/data/Alligators.dat"
ali = read.table(data, header = T)
attach(ali)
y.data = cbind(y2, y3, y4, y5, y1)
x.data = model.matrix(~size + factor(lake), data = ali)
library(VGAM)
# We fit a multinomial logit model with fish (y1) as the reference category:
fit.main = vglm(cbind(y2, y3, y4, y5, y1) ~ size + factor(lake), family = multinomial,
    data = ali)
summary(fit.main)
pchisq(deviance(fit.main), df.residual(fit.main), lower.tail = FALSE)
```

```
##
## Call:
## vglm(formula = cbind(y2, y3, y4, y5, y1) ~ size + factor(lake),
##     family = multinomial, data = ali)
##
##
## Pearson residuals:
##   log(mu[,1]/mu[,5]) log(mu[,2]/mu[,5]) log(mu[,3]/mu[,5])
## 1             0.0953           0.028205           -0.54130
## 2            -0.5082           0.003228            0.66646
## 3            -0.3693          -0.461102           -0.42005
## 4             0.4125           0.249983            0.19772
## 5            -0.5526          -0.191149            0.07215
## 6             0.6500           0.110694           -0.02784
```

```
## 7            0.6757           0.827737           0.79863
## 8           -1.3051          -0.802694          -0.69525
##   log(mu[,4]/mu[,5])
## 1           -0.7268
## 2            1.2589
## 3            1.8347
## 4           -1.3779
## 5            0.2790
## 6           -0.2828
## 7           -0.3081
## 8            0.4629
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1    -3.2074     0.6387  -5.021 5.13e-07 ***
## (Intercept):2    -2.0718     0.7067  -2.931 0.003373 **
## (Intercept):3    -1.3980     0.6085  -2.297 0.021601 *
## (Intercept):4    -1.0781     0.4709  -2.289 0.022061 *
## size:1            1.4582     0.3959   3.683 0.000231 ***
## size:2           -0.3513     0.5800  -0.606 0.544786
## size:3           -0.6307     0.6425  -0.982 0.326296
## size:4            0.3316     0.4482   0.740 0.459506
## factor(lake)2:1   2.5956     0.6597   3.934 8.34e-05 ***
## factor(lake)2:2   1.2161     0.7860   1.547 0.121824
## factor(lake)2:3  -1.3483     1.1635  -1.159 0.246529
## factor(lake)2:4  -0.8205     0.7296  -1.125 0.260713
## factor(lake)3:1   2.7803     0.6712   4.142 3.44e-05 ***
## factor(lake)3:2   1.6925     0.7804   2.169 0.030113 *
## factor(lake)3:3   0.3926     0.7818   0.502 0.615487
## factor(lake)3:4   0.6902     0.5597   1.233 0.217511
## factor(lake)4:1   1.6584     0.6129   2.706 0.006813 **
## factor(lake)4:2  -1.2428     1.1854  -1.048 0.294466
## factor(lake)4:3  -0.6951     0.7813  -0.890 0.373608
## factor(lake)4:4  -0.8262     0.5575  -1.482 0.138378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])
##
## Residual deviance: 17.0798 on 12 degrees of freedom
##
## Log-likelihood: -47.5138 on 12 degrees of freedom
##
## Number of iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):1'
##
## Reference group is level  5  of the response
## [1] 0.1466189
```

Categorical regresion, ordinal model

```r
# Read mental health data from the web:
library(knitr)
data = "http://www.stat.ufl.edu/~aa/glm/data/Mental.dat"
mental = read.table(data, header = T)
library(VGAM)
# We fit a cumulative logit model with main effects of 'ses' and 'life':
fit.imp = vglm(impair ~ life + ses, family = cumulative(parallel = T), data = mental)
# parallell=T gives proportional odds structure - only intercepts differ
summary(fit.imp)
```

```
##
## Call:
## vglm(formula = impair ~ life + ses, family = cumulative(parallel = T),
##     data = mental)
##
##
## Pearson residuals:
##                     Min      1Q  Median     3Q   Max
## logit(P[Y<=1]) -1.568 -0.7048 -0.2102 0.8070 2.713
## logit(P[Y<=2]) -2.328 -0.4666  0.2657 0.6904 1.615
## logit(P[Y<=3]) -3.688  0.1198  0.2039 0.4194 1.892
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -0.2819     0.6231  -0.452  0.65096
## (Intercept):2   1.2128     0.6511   1.863  0.06251 .
## (Intercept):3   2.2094     0.7171   3.081  0.00206 **
## life           -0.3189     0.1194  -2.670  0.00759 **
## ses             1.1112     0.6143   1.809  0.07045 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
## Residual deviance: 99.0979 on 115 degrees of freedom
##
## Log-likelihood: -49.5489 on 115 degrees of freedom
##
## Number of iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##      life       ses
## 0.7269742 3.0380707
```

LMM - random intercept and slope

```r
library(lme4)
fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy)
summary(fm1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 612.09   24.740
##           Days         35.07    5.922   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.825  36.838
## Days          10.467      1.546   6.771
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

GLMM - random intercept and slope Poisson

```r
library("AED")
data(RIKZ)
library(lme4)
fitRI = glmer(Richness ~ NAP + (1 + NAP | Beach), data = RIKZ, family = poisson(link = log))
summary(fitRI)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Richness ~ NAP + (1 + NAP | Beach)
##    Data: RIKZ
##
##      AIC      BIC   logLik deviance df.resid
##    218.7    227.8   -104.4    208.7       40
```

```
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.35846 -0.51129 -0.21846  0.09802  2.45384
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  Beach  (Intercept) 0.2630   0.5128
##         NAP         0.0891   0.2985   0.18
## Number of obs: 45, groups:  Beach, 9
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6942     0.1868   9.071  < 2e-16 ***
## NAP          -0.6074     0.1374  -4.421 9.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##     (Intr)
## NAP 0.121
```

Exam and exam preparation

We take look at the information posted at Blackboard, and the relevant exams are found on the bottom of each module page.

Dates for supervision are also found on Bb.

After TMA4315

What is next in the spring semester?

For the 4th year student

- TMA4250 Spatial statistics
- TMA4268 Statistical learning
- TMA4275 Survival analysis
- TMA4300 Computational statistics
- KLMED8005 Analysis of repeated measurements
- SMED8002 Epidemiology 2
- TDT4300 Datavarehus og datagruvedrift
- TDT4173 Machine learning and case-based reasoning (Big overlap with TMA4268)
- NEVR3004 Neural networks (in the brain)

For the 5th year student

- MA8701 General statistical models Phd course with selected topics relevant for statistical learning and inference.

Also, for the autumn of 2019 the Deep learning course at IDI which up to now was 3.75STP is planned to be an ordinary 7.5STP course.

Course evaluation in TMA4315

Please answer the course evaluation (anonymous): https://kvass.svt.ntnu.no/TakeSurvey.aspx?SurveyID=tma4315h2018

From my heart – I thank all the students for their active participation at the lectures and for being so positive! I really hope this course has given you skills that you will need and use in your future academic life! I look forward to meeting you as `statisticians` in the future! Good luck on your future endeavours!

-Mette