Fall 2006

TMA4315 GENERALIZED LINEAR MODELS

Bo Lindqvist

Department of Mathematical Sciences

NTNU

bo@math.ntnu.no

http://www.math.ntnu.no/~bo/TMA4315/2006h/

Lecturer	Professor Bo Lindqvist, room 1129, Sentralbygg II. Phone (735) 93532. Office hours: To be announced. Email: bo@math.ntnu.no
Exercise lab teacher	Research assistant Sara Martino, room 1126, Sentralbygg II. Phone (735)50221 Office hours: To be announced. Email: martino@math.ntnu.no
Course book	Annette J. Dobson: An introduction to generalized linear models, 2nd ed. Chapman & Hall/CRC, 2002. Here are online data sets and outline solutions to the exercises from the book. Note: The book is available as "E-book" through BIBSYS. In addition, notes/copies about certain topics will be handed out as needed.
Curriculum	Here is PRELIMINARY CURRICULUM (to be updated as time goes).

PRELIMINARY CURRICULUM AND LECTURE PLAN

Main literature is:

Annette J. Dobson: An introduction to generalized linear models, 2nd ed. Chapman & Hall/CRC, 2002.

Last updated August 11, 2006

Week	Topic	Reference to Dobson	Other reference	Comment
34-35	Generel linear model for Normal data. Some repetition from Industrial Statistics course.	Ch. 1-2, 6		
36	Obligatory Exercise 1:			Due early week 37
37-38	General theory of exponential families, GLMs etc.	Ch. 3-5		
39	Binary variables and logistic regression	Ch. 7		
40	Obligatory Exercise 2			Due early week 41
41-42	Nominal and ordinal logistic regression	Ch. 8		
43	Obligatory Exercise 3			Due early week 44
44-45	Count data, Poisson regression	Ch. 9		
46	Obligatory Exercise 4			Due early week 47
47	Overdispersed models, quasi-likelihood models, generalized estimating equations		Handout	

Typical data, multiple regression

Table 6.3 *Carbohydrate, age, relative weight and protein for twenty male insulin-dependent diabetics; for units, see text (data from K. Webb, personal communication).*

Carbohydrate y	Age x_1	Weight x_2	Protein x_3
33	33	100	14
40	47	92	15
37	49	135	18
27	35	144	12
30	46	140	15
43	52	101	15
34	62	95	14
48	23	101	17
30	32	98	15
38	42	105	14
50	31	108	17
51	61	85	19
30	63	130	19
36	40	127	20
41	50	109	15
42	64	107	16
46	56	117	18
24	61	100	13
35	48	118	18
37	28	102	14

Bokmål

Faglig kontakt under eksamen
John Tyssedal tlf. 73593534

EKSAMEN I TMA4315 GENERALISERTE LINEÆRE MODELLER
Fredag 9. desember 2005
Tid: 09.00 – 13.00

Tillatte hjelpemidler: Alle trykte og håndskrevne hjelpemidler. Alle kalkulatorer er tillatt. Sensur: 9. januar 2005.

Oppgave 1

I en studie av hesteskokrabber fant en at hver hunnkrabbe alltid bodde sammen med en hannkrabbe. I tillegg hadde flere hunnkrabber en eller flere hannkrabber like i nærheten, kalt satellitter. Det var av interesse å finne ut hvilke egenskaper ved hunnkrabbene som særlig tiltrakk seg hannkrabber. For hver hunnkrabbe ble det registrert hvor mange satellitter hun hadde, bredden på skallet og fargen på skallet delt inn i kategoriene lyst, middels lyst, middels mørkt og mørkt. Vi innfører nå følgende variabler:

Y = antall satellitter for hver hunnkrabbe

$$Y_1 = \begin{cases} 1 & \textit{dersom en hunnkrabbe har satellitter} \\ 0 & \textit{elles} \end{cases}$$

X_1 = *bredde av skallet*

$$X_2 = \begin{cases} 1 & \textit{dersom skallet er lyst} \\ 0 & \textit{elles} \end{cases}$$

$$X_3 = \begin{cases} 1 & \textit{dersom skallet er middels lyst} \\ 0 & \textit{elles} \end{cases}$$

$$X_4 = \begin{cases} 1 & \textit{dersom skallet er middels mørkt} \\ 0 & \textit{elles} \end{cases}$$

Innsamlet datamateriale for 20 hunnkrabber er gitt nedenfor:

Krabbenr.	y_i	y_{1i}	x_{1i}	x_{2i}	x_{3i}	x_{4i}
1	8	1	28.3	0	1	0
2	4	1	26.0	0	0	1
3	0	0	21.0	0	0	0
4	3	1	25.0	1	0	0
5	8	1	25.7	0	1	0
6	6	1	27.5	0	1	0
7	5	1	26.1	1	0	0
8	4	1	28.9	0	0	1
9	4	1	22.9	0	1	0
10	3	1	26.2	0	0	1
11	8	1	30.0	0	1	0
12	3	1	26.2	0	1	0
13	4	1	25.4	0	1	0
14	0	0	27.5	0	0	0
15	3	1	27.0	0	0	0
16	1	1	24.5	0	1	0
17	1	1	27.3	0	0	1
18	0	0	22.0	0	0	1
19	2	1	30.2	1	0	0
20	3	1	26.0	0	1	0

For å finne ut om bredden på skallet påvirket sannsynligheten for at en hunnkrabbe hadde satellitter ble det utført en logistisk regresjon med Y_1 som responsvariabel. En utskrift med programpakken R er gitt nedenfor. I spørsmålene a), b) og c) vil du ha bruk for denne utskriften.

```
glm(formula = y1 ~ x1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5764   0.1595   0.3649   0.4483   1.0678

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.7637     8.6268  -1.711   0.0870 .
x1            0.6562     0.3518   1.865   0.0621 .
```

En fant det rimelig at antall satellitter for hver hunnkrabbe var Poisson fordelt og en ønsket difor å utføre en Poissonregresjon med log link for å finne ut om det var noen sammenheng mellom forventet antall satellitter og bredden og fargen på skallet til hunnkrabben. En utskrift med programpakken R er gitt nedenfor. I spørsmålene d), e) og f) vil du ha bruk for denne utskriften.

```
glm(formula = y ~ x1 + x2 + x3 + x4, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.71150  -1.07231   0.07435   0.54191   1.50426

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.44922    1.60729  -2.146   0.0319 *
x1           0.13426    0.05737   2.340   0.0193 *
x2           0.96755    0.66514   1.455   0.1458
x3           1.49468    0.59747   2.502   0.0124 *
x4           0.77971    0.64623   1.207   0.2276
```


Example: Birthweight and gestational age

Table 2.3 *Birthweight and gestational age for boys and girls.*

	Boys		Girls	
	Age	Birthweight	Age	Birthweight
	40	2968	40	3317
	38	2795	36	2729
	40	3163	40	2935
	35	2925	38	2754
	36	2625	42	3210
	37	2847	39	2817
	41	3292	40	3126
	40	3473	37	2539
	37	2628	36	2412
	38	3176	38	2991
	40	3421	39	2875
	38	2975	40	3231
Means	38.33	3024.00	38.75	2911.33

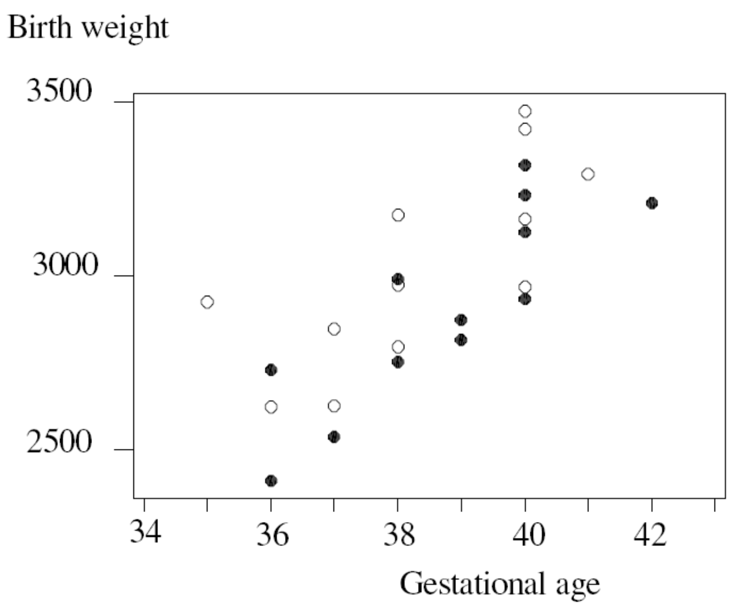

Figure 2.2 *Birthweight plotted against gestational age for boys (open circles) and girls (solid circles); data in Table 2.3.*

Table 2.5 *Analysis of data on birthweight and gestational age in Table 2.3.*

Model	Slopes	Intercepts	Minimum sum of squares
(2.6)	$b = 120.894$	$a_1 = -1610.283$	$\widehat{S}_0 = 658770.8$
		$a_2 = -1773.322$	
(2.7)	$b_1 = 111.983$	$a_1 = -1268.672$	$\widehat{S}_1 = 652424.5$
	$b_2 = 130.400$	$a_2 = -2141.667$	

6

Normal Linear Models

6.1 Introduction

This chapter is about models of the form

$$\mathrm{E}(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} \quad ; \quad Y_i \sim N(\mu_i, \sigma^2) \tag{6.1}$$

where $Y_1, ..., Y_N$ are independent random variables. The link function is the identity function, i.e., $g(\mu_i) = \mu_i$. This model is usually written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{6.2}$$

where

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$$

and the e_i's are independently, identically distributed random variables with $e_i \sim N(0, \sigma^2)$ for $i = 1, ..., N$. Multiple linear regression, analysis of variance (ANOVA) and analysis of covariance (ANCOVA) are all of this form and together are sometimes called **general linear models**.