

Noter til Specialkursus i videregående statistik

Poul Thyregod

IMM, februar 2005

Indhold

Forord	6
1 Momenter og flerdimensionale stokastiske variable	7
1.0 Indledning	7
1.1 Notation	7
1.2 Momenter	8
1.3 Momenter for flerdimensionale variable	10
1.4 Fordelinger afledt af normalfordelingen	14
1.4.1 The log-normal distribution model	14
1.4.2 The χ^2 -distribution model	15
1.4.3 Relation to the normal distribution model	16
1.4.4 The t-distribution model	18
1.4.5 The F-distribution model	21
1.4.6 The Gamma-function	22
1.4.7 Oversigt over en række sædvanlige fordelinger	23
2 Likelihoodfunktion	24
2.0 Introduktion og oversigt	24
2.1 Likelihoodfunktion, Scorefunktion og Informationsmatrix	25
2.2 Eksempler	30
3 Lidt om den generelle lineære model (normalfordelingsteori)	38
3.0 Introduktion og oversigt	38
3.1 Statistisk model, den flerdimensionale normalfordelingsmodel .	39
3.2 Den generelle lineære model	42
3.3 Estimation i en GLM	45

3.4	Fordeling af $\hat{\beta}$, fittede værdier og residualer, estimation af σ^2	48
3.4.1	Fordeling af $\hat{\beta}$, fittede værdier og residualer	48
3.4.2	Kvadratiske former, spaltningssætningen	49
3.4.3	Fordeling af residualkvadratsummen, estimation af σ^2	50
3.4.4	Test af enkelte parametre	51
3.5	Eksempel, simpel lineær regressionsmodel	51
3.6	Reduktion af en generel lineær model	54
3.6.1	Introduktion	54
3.6.2	Kvotienttest for modelreduktion	55
3.6.3	Testet udtrykt i en parametriseret model	60
3.6.4	Modelreduktion ved partielle tests	62
3.6.5	Successiv testning, type I opspaltning	64
3.7	Repræsentation af modeller, modelformler og analytiske udtryk	66
3.7.1	Intercept led	68
3.7.2	Kontinuerte kovariable	69
3.7.3	Kvalitative kovariable, faktorvariable	70
3.7.4	Modelformler	74
3.7.5	Inklusionsdiagrammer for hypotese-kæder, Type I og Type III opspaltning	77
3.8	Eksempel, Blodgennemstrømningsindeks hos malere	81
3.9	Vurdering af modeltilpasning, diagnostiske størrelser	84
3.9.1	Residualer, standardisering og studentisering	84
3.9.2	Kontrol af enkeltobservationer, leverage	86
3.9.3	Kontrol af enkeltobservationers overensstemmelse, residual	89
3.9.4	Kontrol af enkeltobservationers indflydelse (influens)	90
3.10	Ortogonalitet, leverage, Lille eksempel og øvelse	93
3.11	Indskud om Yates' algoritme:	93
4	Introduktion til generaliserede lineære modeller	98
4.0	Oversigt	98
4.1	Baggrund og hovedideer	100
4.2	Motiverende eksempel	101
4.3	Naturlige eksponentielle familier, eksponentielle dispersionsfamilier	104
4.3.1	Naturlige eksponentielle familier	104

4.3.2	Ekspontielle dispersionsfamilier	106
4.3.3	Middelværdiafbildning, kanonisk link, variansfunktion, og enhedsdevians	109
4.4	Model, likelihood og scorefunktion	116
4.4.1	Model for et observationssæt	116
4.4.2	Likelihoodfunktion, scorefunktion og informationsmatrix	117
4.5	Generaliserede lineære modeller	119
4.5.1	Generaliseret lineær model, dimension og modelmatrix	119
4.5.2	Komponenter i en generaliseret lineær model	121
4.5.3	Fuld model	122
4.5.4	Lokal design matrix	123
4.6	Maksimum likelihood estimation	125
4.6.1	Estimation	125
4.6.2	Fordeling af maksimum-likelihood estimat	126
4.6.3	Eksempel på endimensional regression med kanonisk link	128
4.6.4	Fejlophobningsloven, iterativt genvægtet mindste kva- draters metode	133
4.7	Fordeling af fittede værdier og residualer	134
4.7.1	Fittede værdier og residualer	134
4.7.2	Fordeling af fittede værdier og residualer	136
4.7.3	Andre former for residualer	138
4.8	Test for modeltilpasning	141
4.9	Test af enkelte parametre	144
4.10	Test af delhypotese	146
4.11	Successiv testning	151
4.11.1	Eksempel, tofaktor model med binomialt respons . . .	152
4.12	Modeller med overdispersion	161
4.12.1	Quasi-devians og quasi log-likelihood	162
4.13	Værktøjer i SAS og S-plus	163
4.13.1	Generaliserede lineære modeller i SAS Insight	163
4.13.2	Andre SAS-procedurer	165
4.13.3	Generaliserede lineære modeller i S-plus	166
4.14	Eksempler på “regressionsmodeller”	166
4.14.1	Logistisk link ved binomial respons, odds	166
4.14.2	Andre linkfunktioner for binært respons	168
4.14.3	Indskud om sensitivitet, specificitet af klassifikation, ROC-kurve	173

4.14.4	Ordnet kategorisk respons	174
4.14.5	Poisson regression, brug af OFFSET	182
4.14.6	Empiriske varianser for normalfordelte observationer	183
4.14.7	Bartlett's test	188
4.15	Tofaktormodeller, vekselvirkning	188
4.15.1	Grafisk kontrol, profilplot	189
4.15.2	Tofaktormodel for Poissonfordelte data	191
4.15.3	Vekselvirkning og valg af linkfunktion	194
4.16	Linkfunktioner	202
4.17	Lidt mere om eksponentielle familier	203
5	Normalfordelingsmodeller med hierarkisk variation	213
5.1	Betingede, simultane og marginale fordelinger	214
5.1.1	Lidt om betingede og marginale middelværdier og varianser	216
5.2	Envejsanalyse, tilfældig model	217
5.2.1	Eksempel	217
5.3	Model	218
5.3.1	Marginale og simultane fordelinger	219
5.4	Test af homogenitetshypotese	222
5.5	Estimation af faste parametre	224
5.6	Estimation af de tilfældige effekter, BLUP-estimation	229
5.7	Lidt mere om hierarkiske modeller	230
5.7.1	Generelt	230
5.7.2	Aposteriorifordeling svarende til normal-normal fordeling	232
5.8	I SAS	235
5.8.1	Mixed	239
5.9	Empirisk Bayes estimation, BLUP estimation	243
5.10	Flere eksempler på hierarkisk variation for normalfordelingsmodeller	249
5.10.1	Analyse af blokforsøg	250
5.10.2	Varierende regressionslinier	251
5.10.3	Flerdimensional normalfordeling	255
5.11	Aposteriorifordelinger for flerdimensionale normalfordelinger	259
6	Generelle lineære mixede modeller	261
6.0	Introduktion	261
6.1	Formulering af den generelle lineære mixed model	261

6.1.1	Eksempel: Envejs model med tilfældig effekt	262
6.2	Estimation	263
6.2.1	Estimation af fixed effects og variansparametre	263
6.2.2	Estimation af de tilfældige effekter	264
6.2.3	Samtidig estimation af β og $\mathbf{u}$	265
6.2.4	Empirisk Bayes-fortolkning	266
6.3	SAS-proceduren PROC MIXED	267
6.3.1	Eksempel på brug af PROC MIXED	267
7	Hierarkisk variation for exponentielle dispersionsfamilier	270
7.0	Indledning, modellering af overdispersion	270
7.1	Hierarkisk Poisson Gamma model	271
7.1.1	Eksempel, Variation mellem episoder af tordenvejr ved Cape Kennedy	271
7.1.2	Formulering af hierarkisk model	272
7.1.3	SAS PROC GENMOD	275
7.1.4	Eksempel, tordenvejr igen	275
7.2	Hierarkisk binomial beta model	276
7.2.1	Data	276
7.2.2	Formulering af hierarkisk model	278
7.3	Hierarkisk model for empiriske varianser for normaltfordelte variable	279
7.3.1	Den systematiske model	279
7.3.2	Den tilfældige model	280
7.3.3	Fortolkning af parametre i strukturfordelingen af σ^2	281
7.3.4	Marginal fordeling af stikprøvevariansen	282
7.4	Normalfordelingsmodeller med tilfældigt varierende varians.	286
7.5	Generel formulering af hierarkiske modeller for exponentielle dispersionsfamilier	287
7.6	Aposteriorifordeling	290
7.6.1	Generelle resultater vedrørende aposteriorifordelinger	291
7.7	Aposteriorifordeling for Poisson Gamma model	296
7.8	Normalfordelingen som miksturfordeling	297
8	sasdata	301
9	Diverse	307
9.1	om information	307