ЕМ алгоритми за *"probit"* модели със случайни ефекти

Деница Григорова

Катедра ВОИС

Факултет по Математика и Информатика

Софийски Университет "Св. Климент Охридски"

Автореферат към дисертация, предадена за придобиване на образователна и научна степен

Доктор

Декември 2013

Научен ръководител:

доц. д-р Ралица Георгиева

Резюме

В дисертацията са разработени ЕМ алгоритми за три различни "probit" модела със случайни ефекти. Алгоритмите са теоретично описани и реализирани в безплатната среда за статистическа обработка на данни **R** (R Core Team [2013]). Два от моделите са за данни с повторни наблюдения над обектите. Единият от тях е за една наредена категорна променлива с повторни наблюдения. Другият е съвместен модел на две, проследени неколкократно във времето, величини от различен тип. Предполагаме, че едната е нормално разпределена, а другата - наредена категорна. Третият ЕМ алгоритъм е за съвместен модел за няколко наредени категорни променливи. За всички ЕМ алгоритми са направени симулационни изследвания с цел да се провери асимптотична неизместеност на оценките. С помощта на алгоритмите са оценени параметрите на статистически модели, приложени върху реални данни от областта на медицината.

Съдържание

Глава 1

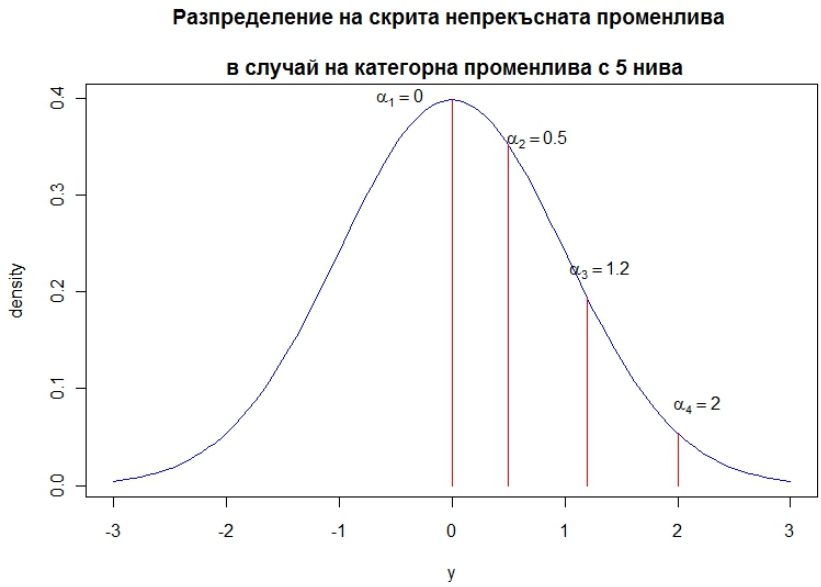
Въведение

1.1 Литературен обзор

За статистическото моделиране на една променлива, проследявана неколкократно във времето, има три различни широко разпространени групи статистически модели. Това са модели със случайни ефекти, маргинални модели и транзитивни модели. Всички те отчитат корелацията на наблюденията, която съществува в рамките на един обект. Книгата на Fitzmaurice et al. [2004] дава изключително изчерпателен поглед върху моделите, методите за оценяване и проблемите при работа с данни с повторни наблюдения.

Един подход за моделирането на наредена категорна величина е допускането за ненаблюдавана латентна непрекъсната променлива, която поражда наблюдаваната. Предполага се, че непрекъснатата променлива може да се раздели на интервали, чиито брой е равен на броя на нивата на категорната променлива. Когато латентната случайна величина приеме дадена стойност, то качествената приема това ниво, за което отговаря интервалът, в който попада непрекъснатата (пример е Фигура 1.1). Такива модели са наречени модели с праг и в книгата на Fahrmeir and Tutz [2001] са разгледани такива модели за бинарни и наредени категорни данни.

За пръв път "probit" моделите са предложени от Gaddum [1933] и Bliss [1934a,b] за бинарни данни. Ashford and Sowden [1970] представят многомерно разширение на "probit" модел, базирано на скрито многомерно нормално



Фигура 1.1: Нормално разпределена скрита променлива с 4 прага, която генерира наблюдавана категорна променлива с 5 нива.

разпределение. Разширения на тези модели са предложени от Gibbons and Hedeker [1994], Catalano [1997], Grilli and Rampichini [2003], Gueorguieva and Sanacora [2006], Kawakatsu and Largey [2009], както и други. Корелираният "probit" модел се използва широко, защото е лесен за интерпретация и позволява различни корелационни структури в рамките на обекта. Gueorguieva [2006] има детайлна статия върху корелираните "probit" модели. Проблемите с намирането на оценките на неизвестните параметри на тези модели е все още област, в която се работи интензивно.

Корелираните "probit" модели нямат явен вид на функцията на правдоподобие и е необходимо да се използват апроксимации за оценяването на параметрите. Има развити няколко подхода за общия случай за модели за зависими променливи от експоненциалното семейство (McCullagh and Nelder [1989]), към които спадат Бернулиево, нормално, Поасоново, гама разпределения. Това са числени, стохастични и аналитични приближения.

Най-често се използват разширения на числените апроксимации като Гаус-Хермитова квадратура (описана на стр. 306-307 в Fahrmeir and Tutz [2001])

или адаптирана Гаусова квадратура (Liu and Pierce [1994]). Тези подходи имат недостатък да изискват голяма изчислителна мощност, когато размерността на случайните ефекти нараства. Такива методи са разработени и използвани от Gueorguieva and Sanacora [2006] и Grilli and Rampichini [2003].

Друг подход е аналитично приближение (Breslow and Clayton [1993], Wolfinger and O'Connell [1993]), но е показано, че оценките на параметрите са изместени за бинарни данни и ординални данни с малко на брой категории.

Стохастичните приближения (McCulloch [2008], описани на стр. 311-315 в Fahrmeir and Tutz [2001]) се оказват най-удачни за оценяване на параметрите на модели със случайни ефекти, тъй като изчислителната сложност не нараства експоненциално с нарастването на броя на случайните ефекти и дават асимптотично неизместени оценки. Такива методи чрез EM алгоритми (Dempster et al. [1977]) са разработени за корелирани бинарни и нормални данни от Gueorguieva and Agresti [2001], но не са разработени за корелирани ординални данни.

Алтернативни методи за намиране на оценки на параметрите на корелирани "probit" модели са чрез Бейсови подходи (Dunson et al. [2003]) и методи на обобщените уравнения за оценяване (от англ. Generalized Estimating Equations, Geys et al. [2001]).

Дисертацията включва развиването на ECM алгоритми за оценяването на неизвестните параметри на три корелирани "probit" модела. Първият е модел за една наредена категорна променлива с повторни наблюдения. Следващият е корелиран "probit" модел за няколко наредени категорни променливи. Третият е съвместен модел за две променливи, проследени неколкократно във времето от различен тип - нормална и наредена категорна. За всички модели ще предполагаме, че зад наблюдаваната категорна променлива има цензурирана латентна непрекъсната променлива. За отчитане на корелацията между повторните наблюдения над величините от основен интерес ще въведем случайни ефекти. Методите разширяват метода за оценяване, разработен от Kawakatsu and Largey [2009], за една наредена категорна величина и многомерен нормален отклик, в случаите за корелирани наредени категорни променливи изброени по-горе.

В приложенията на дисертацията са представени функциите за реализи-

рането на ECM алгоритмите в свободната и широко разпространена в академичната общност среда за статистическа обработка на данни **R** (R Core Team [2013]).

1.2 Структура на дисертацията

В Главата Модел за една наредена категорна променлива с повторни наблюдения е предложен "probit" модел за една наредена категорна променлива с повторни наблюдения. Развит е ECM алгоритъм за максимално правдоподобни оценки (МПО) на неизвестните параметри в изложения модел. Представени са симулации с цел изследване на свойствата на алгоритъма.

Главата Съвместен модел за няколко наредени категорни променливи представя модел за многомерни наредени категорни данни. Предложен е ECM алгоритъм за МПО на неизвестните параметри в модела. За реализацията на алгоритъма е създадена функция, с помощта на която са извършени симулации, за да се изследва поведението на алгоритъма. Чрез алгоритъма е оценен "probit" модел за изследване на връзката между генотипа и степента на два типа (кожни и урогенитални) странични реакции след лъчетерапия при онкоболни жени.

Главата Съвместен модел за една наредена категорна и една нормална променливи, проследени неколкократно във времето е за съвместен модел за една категорна и една нормална променливи с повторни наблюдения. Описан е ECM алгоритъм за МПО на неизвестните параметри. За да се потвърди надеждността на предложения алгоритъм, са представени две симулационни изследвания.

В главата Заключение са изложени основните приноси на дисертацията и бъдещи насоки за развитие на представените резултати.

Глава 2

Модел за една наредена категорна променлива с повторни наблюдения

В настоящата глава е разгледан корелиран "probit" модел със случайни ефекти за една наредена категорна променлива от дългосрочни проучвания. За намиране на МПО на неизвестните параметри в модела е предложен ECM алгоритъм. Две симулационни изследвания потвърждават надеждността на алгоритъма. Чрез него е оценен корелиран "probit" модел, приложен към реални данни от американско дългосрочно проучване. Представените по-долу резултати са публикувани в Grigorova and Gueorguieva [2013a].

2.1 Описание на модела

Нека с y_{ij}^* означим наблюдението над i-ти обект във време j над наредена категорна променлива с m нива ($i = 1, \ldots, n,\ j = 1, \ldots, n_i$). Ще предполагаме, че съществува скрита нормално разпределена променлива y_{ij}, която поражда наблюдаваната величина. Разглеждаме следния модел със случайни ефекти за скритата променлива:

$$y_{ij} \quad = \quad \boldsymbol{x'_{ij}\beta} + \boldsymbol{z'_{ij}b_i} + \epsilon_{ij}. \tag{2.1}$$

Правилото, което свързва скритата непрекъсната с наблюдаваната категорна величина, е:

$$y_{ij}^{*} = \begin{cases} 1, \ y_{ij} \leq \alpha_1; \\ l, \ \alpha_{l-1} < y_{ij} \leq \alpha_l, \ l = 2, \ldots, m-1; \\ m, \ y_{ij} > \alpha_{m-1}; \end{cases} \tag{2.2}$$

за неизвестни прагове $\alpha_1, \ldots, \alpha_{m-1}$.

В модела 2.1 предполагаме, че векторът от случайните ефекти е нормално разпределен с размерност q и сме го означили с $\boldsymbol{b_i} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Ковариационната матрица на случайните ефекти $\boldsymbol{\Sigma}$ е квадратна матрица с размерност $q \times q$ и е положително полу-дефинитна. Допускаме също, че грешките са нормално разпределени $\epsilon_{ij} \sim N(0, \sigma^2)$, независими помежду си и независими от случайните ефекти.

Регресионните параметри за фиксираните ефекти са означени с p-мерния вектор $\boldsymbol{\beta}$. Векторът с предсказващите променливи за фиксираните ефекти е $\boldsymbol{x_{ij}}$ и векторът с предсказващите променливи за случайните ефекти е $\boldsymbol{z_{ij}}$.

Нека с вектора $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{in_i})', i = 1, \ldots, n$ е означен набора от всички ненаблюдавани непрекъснати величини над обекта i, с вектора $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)'$ всички случайни ефекти и с вектора $\boldsymbol{y} = (\boldsymbol{y_1'}, \boldsymbol{y_2'}, \ldots, \boldsymbol{y_n'})'$ набора от всички ненаблюдавани скрити величини над всички обекти.

От наблюдаваните данни не могат да бъдат еднозначно оценени всички неизвестни параметри в модела 2.1 и неизвестните прагове в 2.2, затова налагаме идентификационни ограничения: първият праг α_1 е фиксиран в нулата, а дисперсията на нормалните грешки σ^2 е 1. Възможни са други параметрични ограничения и репараметризации за модела с цел еднозначна определеност. Под еднозначна определеност се има предвид, че при зададен модел това води до единствена възможност за стойностите на параметрите.

2.2 Намиране на максимално правдоподобни оценки (МПО) чрез ЕМ алгоритъм

Ние правим разширение на стохастичния ECM алгоритъм, предложен от Chan and Kuk [1997], за да оценим неизвестните параметри в модела 2.1 и неизвестните прагове в 2.2. Първата стъпка е въвеждането на праговете във функцията на пълното правдоподобие. За тази цел използваме подхода на Kawakatsu and Largey [2009], които разширяват работата на Ruud [1991]. Дефинираме разликите между съседните прагове с $\delta_i = \alpha_i - \alpha_{i-1},\ i = 2, \ldots, m-1$. Оттук следва връзката, че $\alpha_i = \sum_{k=2}^{i} \delta_k,\ i = 2, \ldots, m-1$. За пълнота и по-нататъшна употреба определяме $\delta_1 = \delta_m = 1$. Следващата стъпка е да разгледаме нова променлива $y_{ij_{new}}$, която е линейна трансформация на скритата величина y_{ij}: $y_{ij_{new}} = (y_{ij} - \alpha_{y^*_{ij}-1})/\delta_{y^*_{ij}}$. Отново с цел пълнота дефинираме $\alpha_0 = 0$. Понеже новата променлива е линейна трансформация на нормална сл. вел., то тя също има нормално разпределение. Тази променлива, условно по наблюдаваната категорна променлива, има орязано нормално разпределение и от дефиницията ѝ следва, че границите на орязване не зависят от неизвестните параметри. Ако наблюдаваме първото ниво на y^*, то трансформираната променлива е орязана в интервала $(-\infty, 0]$. Ако y^* е между първото и последното ниво, орязването е в интервала $(0, 1]$. Ако сме наблюдавали последното ниво на y^*, новата променлива е орязана на $(0, \infty)$.

2.2.1 Лог-правдоподобие на пълния набор от данни

Пълният набор от данни представлява данните, които бихме събрали, ако бяхме наблюдавали скритата непрекъсната величина и случайните ефекти. Пълното лог-правдоподобие $\ln L$ е плътността на многомерното разпределение на случайните ефекти и линейните трансформации на скритите нормални величини в стойностите на пълния набор от данни:

$$\ln L = \ln f(\boldsymbol{b}, \boldsymbol{y_{new}}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{i_{new}}} | \boldsymbol{b_i}) = \sum_{i=1}^{n} \ln[f(\boldsymbol{b_i}) \prod_{j=1}^{n} f(y_{ij_{new}} | \boldsymbol{b_i})],$$

където $\boldsymbol{y_{new}} = (\boldsymbol{y'_{1_{new}}}, \boldsymbol{y'_{2_{new}}}, \ldots, \boldsymbol{y'_{n_{new}}})'$ и $\boldsymbol{y_{i_{new}}} = (y_{i1_{new}}, y_{i2_{new}}, \ldots, y_{in_{i_{new}}})', i = 1, \ldots, n.$

Като изключим константите, пълното лог-правдоподобие има следния явен вид:

$$\begin{aligned} \ln L &= -0.5 \sum_{i=1}^{n} \ln |\boldsymbol{\Sigma}| - 0.5 \sum_{i=1}^{n} \boldsymbol{b'_i} \boldsymbol{\Sigma^{-1}} \boldsymbol{b_i} + \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{n_i} \ln \delta_{y^*_{ij}} - 0.5 \sum_{i=1}^{n} \sum_{j=1}^{n_i} [\delta_{y^*_{ij}} y_{ij_{new}} - (\boldsymbol{x'_{ij}} \boldsymbol{\beta} + \boldsymbol{z'_{ij}} \boldsymbol{b_i} - \alpha_{y^*_{ij}-1})]^2. \end{aligned}$$

Диференцирайки лог-правдоподобието и приравнявайки първите производни на нула, ние получаваме затворени форми за оценките на неизвестните параметри $\boldsymbol{\Gamma} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\delta})$, където $\boldsymbol{\delta} = (\delta_2, \ldots, \delta_{m-1})$.

2.2.2 Явен вид на МПО

Оценката за ковариационната матрица на случайните ефекти $\boldsymbol{\Sigma}$ за пълните данни е:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=0}^{n} \boldsymbol{b_i} \boldsymbol{b'_i}.$$

Регресионните параметри за фиксираните ефекти удовлетворяват следната система от уравнения:

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \boldsymbol{x_{ij}} \boldsymbol{x'_{ij}} \boldsymbol{\beta} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} [\delta_{y^*_{ij}} y_{ij_{new}} - \boldsymbol{z'_{ij}} \boldsymbol{b_i} + \alpha_{y^*_{ij}-1}] \boldsymbol{x_{ij}}.$$

Уравненията за $\delta_k,\ k = 2, \ldots, m-1$ са квадратни от вида: $a\delta_k^2 + b\delta_k + c = 0$, където константите a, b, c са:

$$\begin{aligned} a &= \sum_{i,j} \sum_{y^*_{ij}=k} (y^2_{ij_{new}}) + n_{k+1} + \ldots + n_m, \\ b &= -\sum_{i,j} \sum_{y^*_{ij}=k} y_{ij_{new}} (\boldsymbol{x'_{ij}} \boldsymbol{\beta} + \boldsymbol{z'_{ij}} \boldsymbol{b_i} - \alpha_{k-1}) + \end{aligned}$$

$$\sum_{i,j}\sum_{y^*_{ij}>k}(\delta_{y^*_{ij}}y_{ij_{new}} - \boldsymbol{x'_{ij}\beta} - \boldsymbol{z'_{ij}b_i} + \delta_2 + \ldots + \delta_{k-1} + \delta_{k+1} + \ldots + \delta_{y^*_{ij}-1}),$$
$$c = -n_k,$$

и n_k е броят на наблюденията на категорната величина на $k-$то ниво. Уравненията винаги имат реални корени и по-големият от тях е положителен.

За да обновяваме оценките на всяка стъпка на алгоритъма, трябва да изчисляваме очакванията на сл.вел., участващи в описаните по-горе оценки при условие наблюдаваните данни. Тези условни математически очаквания зависят само от първите два момента на многомерно орязано нормално разпределение.

2.2.3 Условни очаквания

Нека използваме следните означения за опростяване на записа:

$$\boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{x'_{i1}} \\ \boldsymbol{x'_{i2}} \\ \vdots \\ \boldsymbol{x'_{in_i}} \end{pmatrix}, \boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{z'_{i1}} \\ \boldsymbol{z'_{i2}} \\ \vdots \\ \boldsymbol{z'_{in_i}} \end{pmatrix}, \boldsymbol{\alpha_i} = \begin{pmatrix} \alpha_{y^*_{i1}-1} \\ \alpha_{y^*_{i2}-1} \\ \vdots \\ \alpha_{y^*_{in_i}-1} \end{pmatrix}, \boldsymbol{\delta_i^{-1}} = \begin{pmatrix} 1/\delta_{y_{i1}} \\ 1/\delta_{y_{i2}} \\ \vdots \\ 1/\delta_{y_{in_i}} \end{pmatrix}.$$

Тогава съвместното разпределение на $\boldsymbol{y_{i_{new}}}$ и $\boldsymbol{b_i}$ е многомерно нормално:

$$\begin{pmatrix} \boldsymbol{y_{i_{new}}} \\ \boldsymbol{b_i} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{(X_i\beta - \alpha_i)} \circ \boldsymbol{\delta_i^{-1}} \\ \boldsymbol{0} \end{pmatrix}, \boldsymbol{V}\right],$$

където $\circ$ е поелементно умножение (умножение на Адамар), ковариационната матрица $\boldsymbol{V}$ е:

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{(Z_i\Sigma Z'_i + I_{n_i})} \circ \boldsymbol{\delta_i^{-1}\delta_i^{-1'}} & \boldsymbol{Z_i\Sigma} \circ \boldsymbol{(J_{n_i\times q}\delta_i^{-1})} \\ \boldsymbol{\Sigma Z'_i} \circ \boldsymbol{(J_{n_i\times q}\delta_i^{-1})'} & \boldsymbol{\Sigma} \end{pmatrix}$$

и $\boldsymbol{J_{n_i\times q}\delta_i^{-1}}$ е $n_i \times q$ матрица с колони $\boldsymbol{\delta_i^{-1}}$.

Нека да означим с

$$\boldsymbol{M_i} = \boldsymbol{y_{i_{new}}} - \boldsymbol{(X_i\beta - \alpha_i)} \circ \boldsymbol{\delta_i^{-1}} \text{ и със}$$
$$\boldsymbol{\Sigma_{B_i}} = [\boldsymbol{\Sigma Z'_i} \circ \boldsymbol{(J_{n_i\times q}\delta_i^{-1})'}][\boldsymbol{(Z_i\Sigma Z'_i + I_{n_i})} \circ \boldsymbol{\delta_i^{-1}\delta_i^{-1'}}]^{-1}.$$

Тогава условното разпределение на $\boldsymbol{b_i}$ при условие $\boldsymbol{y_{i_{new}}}$ е отново нормално:

$$\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}} \sim N[\boldsymbol{\Sigma_{B_i}} \boldsymbol{M_i}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma_{B_i}}(\boldsymbol{Z_i}\boldsymbol{\Sigma} \circ (\boldsymbol{J_{n_i \times q}}\boldsymbol{\delta_i^{-1}}))] \ .$$

В изразите на оценките на Е-стъпката на алгоритъма участват следните условни очаквания: $E(\boldsymbol{b_i}|\boldsymbol{y_i^*}), E(\boldsymbol{b_i}\boldsymbol{b_i'}|\boldsymbol{y_i^*}), E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_i^*})$. В дисертацията е показано, че тези условни очаквания зависят само от първите два момента на $\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}$, което е многомерно орязано нормално разпределение.

Условните очаквания е възможно да бъдат изчислени с аналитични форми според работата на Manjunath and Wilhelm [2009], но използваната апроксимация за реализацията на алгоритъма е стохастична и дава стойности достатъчно близки до истинските. За да намерим условните очаквания, е използван Монте Карло метод. Той се състои в генериране на числа от многомерно орязано нормално разпределение при условие наблюдаваните данни. За реализацията е използван методът за генериране на числа на Гибс (от англ. Gibbs sampling, Casella and George [1992]).

Използваме извадъчните средно и дисперсия на получената по този метод извадка, за да намерим приближени стойности на $E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*})$ и $Var(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*})$.

2.2.4 $(k+1)$-ва итерация на ECM алгоритъма

Ние използваме разширение на EM алгоритъма, наречено условно максимизиране в М-стъпката (Expectation/Conditional Maximisation algorithm (Meng and Rubin [1993])). Разликата на ECM алгоритъма от EM алгоритъма е, че в М-стъпката всеки параметър се максимизира индивидуално при условие, че останалите параметри стоят фиксирани. М-стъпката се състои от толкова под-стъпки, колкото брой неизвестни параметри имаме. Оценките на параметрите се изчисляват в една и съща последователност на всяка итерация. За обновяването на конкретен параметър се използват пресметнатите до момента стойности на останалите параметри (включително обновените в настоящата итерация параметри).

Оценките на неизвестните параметри на $k+1-$вата стъпка $\boldsymbol{\Gamma^{k+1}}$ на предлагания ECM алгоритъм са следните:

- $(k+1)$-вата оценка на регресионните параметри $\boldsymbol{\beta^{k+1}}$ са решение по

МНМК при регресия на $E(\tilde{y}_{ij}|\boldsymbol{y_i^*};\boldsymbol{\Gamma^k})$ по $\boldsymbol{x_{ij}}$.

- $(k+1)$-вата оценка за $\delta_u,\ u=2,\ldots,m-1$ е: $\delta_u^{k+1} = (-E[b|\boldsymbol{y}^*;\boldsymbol{\Gamma^k}] + \sqrt{(E[b|\boldsymbol{y}^*;\boldsymbol{\Gamma^k}]^2 - 4E[a|\boldsymbol{y}^*;\boldsymbol{\Gamma^k}]E[c|\boldsymbol{y}^*;\boldsymbol{\Gamma^k}])})/2E[a|\boldsymbol{y}^*;\boldsymbol{\Gamma^k}]$.

  За изчисляването на очакванията на изразите за a, b, c се използват обновените оценки $\boldsymbol{\beta^{k+1}}$, $\delta_i^{k+1}, i=2,\ldots,u-1$.

- $(k+1)$-вата оценка на ковариационната матрица на случайните ефекти е $\hat{\boldsymbol{\Sigma}}^{k+1} = \frac{1}{n}\sum_{i=0}^{n} E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*};\boldsymbol{\Gamma^k})$, където за изчисленията на очакванията използваме обновените оценки $\boldsymbol{\beta^{k+1}}$, $\delta_i^{k+1}, i=2,\ldots,m-1$.

За да пресметнем оценките, използваме приближени стойности за очакванията и дисперсиите.

Алгоритъмът спира, когато разликите между стойностите на параметрите в две съседни итерации станат по-малки от предварително зададено малко число ϵ (например $\epsilon = 0.0001$). Ако искаме да постигнем по-голяма точност за намерените оценки, трябва да използваме по-малка стойност за ϵ.

2.2.5 Приближение на стандартните грешки

Ние използваме "bootstrap" метод, описан в McLachlan and Krishnan [2008] стр. 130-131, за да намерим приближение за стандартните грешки на оценките. "Bootstrap" методът попада в рамката на по-широк клас методи за повторни извадки. Използва се в случаите, когато е неизвестно разпределението на статистиката от интерес (например при тестване на хипотези, построяване на доверителни интервали или в разглеждания случай за намиране на приближение на стандартни грешки). Предимство на метода е неговата простота. При сложни модели изисква време, т.к. броят на изчисленията е голям.

2.3 Симулации

Симулирани са стойности от следния модел със случаен свободен член:

$$y_{ij} \quad = \quad \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij},\ j=1,\ldots,5, \tag{2.3}$$

където $\beta_0 = -0.5, \beta_1 = 1, Var(b_i) = \sigma_b = 0.01, Var(\epsilon_{ij}) = 1$ с прагове $\alpha_1 = 0, \alpha_2 = 1.5, \alpha_3 = 3, \alpha_4 = 4$.

Симулирахме 100 извадки с различен брой индивиди в извадката ($n = 100$ и $n = 500$) с по 5 наблюдения над обект. За всяко приближение на стандартните грешки сме използвали 75 "bootstrap" извадки, което е в рамките на препоръчителния брой между 50 и 100 "bootstrap" повторения (Efron and Tibshirani [1994]). Резултатите са представени в Таблица 2.1.

Таблица 2.1: Оценки и стандартни грешки от двете симулационни изследвания за "probit" модела 2.3 за повторни наблюдения над една наредена категорна променлива

истински стойности	$\beta_0 = -0.5$	$\beta_1 = 1$	$\delta_2 = 1.5$	$\delta_3 = 1.5$	$\delta_4 = 1$	$\sigma_b = 0.01$
Симулация 1: брой на обектите = 100, $n_i = 5$						
средно на оценките	-0.498	1.007	1.512	1.498	1.009	0.011
стд. откл. на оценките	0.114	0.051	0.107	0.108	0.090	0.010
средно на "bootstrap" стд. гр.	0.138	0.052	0.122	0.112	0.094	0.012
Симулация 2: брой на обектите = 500, $n_i = 5$						
средно на оценките	-0.503	1.001	1.50	1.50	0.997	0.010
стд. откл. на оценките	0.06	0.023	0.053	0.052	0.036	0.0006
средно на "bootstrap" стд. гр.	0.059	0.023	0.052	0.049	0.040	0.0006

Да отбележим, че поради репараметризацията, ние оценяваме разликите в праговете, а не самите тях. И за двете симулации средните на оценките на параметрите са равни на истинските стойности на параметрите в рамките на две цифри след десетичната точка. Изключение прави само последната разлика в праговете при първата симулация с по-малък брой обекти в изследването, но разликата е в рамките на 0.01. Така емпирично установяваме неизместеност на оценките, получени от приложението на алгоритъма.

От наблюдението, че стандартните отклонения на оценките и "bootstrap" стандартните грешки са много близки по стойност, можем да заключим, че алгоритъмът схожда, както се очаква.

2.4 Приложение на модела

Прилагаме разгледания "probit" модел към данните от Здравно и Пенсионно Проучване (ЗПП, http://hrsonline.isr.umich.edu/), проведено сред американски пенсионери и техните брачни партньори. На всеки две години участниците в изследването са запитвани за тяхната самооценка за здраве, дали пушат, дали пият, какъв е индексът им на телесна маса, както и други характеристики, които имат отношение към здравето на индивидите. В настоящия раздел ще изследваме как самооценката за здравето се променя във времето. Променливата от основен интерес е наредена категорна. Тази променлива взема стойности от отлично здраве (1) до лошо (5). Категории (2), (3) и (4) означават много добро, добро и задоволително здраве според самооценката на всеки индивид. Прилагаме следния корелиран "probit" модел със свободен случаен член към данните:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{ij}, \\ \text{srh}_{ij} &= \begin{cases} 1, \ y_{ij} \leq \alpha_1 = 0, \\ l, \ \alpha_{l-1} < y_{ij} \leq \alpha_l, \ l = 2,3,4, \\ 5, \ y_{ij} > \alpha_4, \end{cases} \end{aligned} \tag{2.4}$$

където $\epsilon_{ij} \sim N(0,1), j = 1,\ldots,7$, $Var(b_i) = \sigma_b$ и с 'srh_{ij}' сме отбелязали скритата самооценка за здравето на $i-$ти индивид в момент от време j.

В анализа са включени 7550 индивида от изследването, които имат пълен набор от наблюдения. Резултатите са представени в Таблица 2.2.

Таблица 2.2 показва, че всички параметри в модела са статистически значимо отличими от нула. Параметърът от най-голям интерес е регресионният коефициент β_1. Той е положителен и стойността на статистиката от статистическия тест, че регресионният коефициент β_1 е равен на нула, е $z = 0.12/0.0026 = 46.15$ с вероятностна стойност $p - value < 0.0001$. Следо-

Таблица 2.2: Таблица с оценки и стандартни грешки от модела 2.4, приложен към данните от здравно и пенсионно проучване

	β_0	β_1	δ_2	δ_3	δ_4	σ_b
оценки	1.230	0.115	1.581	1.499	1.418	2.150
стд. гр.	0.015	0.0026	0.011	0.011	0.017	0.049
z-score	80.38	44.57	146.63	142.17	83.80	43.51

вателно според данните можем да заключим, че самооценката за здравето се занижава с времето. Резултатите са очаквани, т.к. здравето на индивидите в пенсионна възраст се влошава. Допълнително изследване може да разкрие дали дадени характеристики като пушене, пиене, спортуване са също свързани със самооценката за здравето.

Също така се забелязва, че дисперсията на случайните ефекти е статистически значимо различна от 0 (z-score$_{\sigma_b}$=43.51, χ^2 =z-score$^2_{\sigma_b}$ = 43.51^2 = 1893.12). Това предполага, че има различия между обектите в самооценките за здравето и има силна корелация между наблюденията над един индивид. Корелацията между латентните величини в рамките на индивида се нарича "polychoric" корелация (Drasgow [2004]) и нейната оценка за тези данни е 0.68.

2.5 Заключение

В тази глава от дисертацията е разгледан корелиран "probit" модел със случайни ефекти за наредени категорни величини от дългосрочни проучвания. Предложено е разширение на ECM алгоритъма на Chan and Kuk [1997] за получаване на МПО, което е реализирано в безплатната среда за статистическия изчисления и графики **R** (R Core Team [2013]). Изследвана е надеждността му чрез симулации. Чрез предложения алгоритъм е оценен корелиран "probit" модел за данни от ЗПП.

Глава 3

Съвместен модел за няколко наредени категорни променливи

В тази глава е разгледан съвместен "probit" модел за няколко наредени категорни променливи. За намиране на МПО на неизвестните му параметри е предложено разширение на ECM алгоритъма, представен в предишната глава. Части от изложените по-долу резултати са публикувани в изданието към Докторантска Конференцията по Математика, Информатика и Обучение, 2013 (http://mie.uni-sofia.bg/, Grigorova and Gueorguieva [2013b]). Цялостните резултатите са приети за публикация в Serdica Journal of Computing през декември 2013.

3.1 Описание на модела

Нека са наблюдавани p категорни променливи $y_j^*, j = 1, \ldots, p$ с наредба на нивата. Предполагаме, че броят на нивата на категориите на величините са съответно $m_j, j = 1, \ldots, p$. Наблюденията над $i-$ти обект са отбелязвани с $\boldsymbol{y_i^*} = (y_{i1}^*, y_{i2}^*, \ldots, y_{ip}^*)', i = 1, \ldots, n$. Предполагаме, че има скрити нормални величини $y_{ij},\ j = 1, \ldots, p$, които генерират наблюдаваните категорни променливи. За да отчетем възможна корелация между категорните величини, разглеждаме следния модел със случайни ефекти за скритите променливи:

$$y_{ij} \;=\; \boldsymbol{x'_{ij}\beta_j} + \boldsymbol{z'_{ij}b_{ij}} + \epsilon_{ij},\ j = 1, \ldots, p, \quad \text{където сме наблюдавали} \tag{3.1}$$

$$y_{ij}^* = \begin{cases} 1, \ y_{ij} \leq \alpha_{j,1}; \\ l, \ \alpha_{j,l-1} < y_{ij} \leq \alpha_{j,l}, \ l = 2, \ldots, m_j - 1; \\ m_j, \ y_{ij} > \alpha_{j,m_j-1}; \end{cases} \tag{3.2}$$

за неизвестни прагове $\alpha_{j,1}, \ldots, \alpha_{j,m_j-1}, \ j = 1, \ldots, p$.

Предполагаме, че векторът от случайните ефекти $\boldsymbol{b_i} = (\boldsymbol{b_{i1}}, \boldsymbol{b_{i2}}, \ldots, \boldsymbol{b_{ip}})'$ е нормално разпределен $N(\mathbf{0}, \boldsymbol{\Sigma})$ с размерност q. Ковариационната матрица на случайните ефекти $\boldsymbol{\Sigma}$ е квадратна матрица с размерност $q \times q$ и е положително полу-дефинитна. Допускаме също, че грешките са нормално разпределени $\epsilon_{ij} \sim N(0, \sigma^2)$, независими помежду си и независими от случайните ефекти.

Регресионните параметри за фиксираните ефекти в модел 3.1 означаваме с q_j-мерните вектори $\boldsymbol{\beta_j}, j = 1, \ldots, p$. Векторите с предсказващите променливи за фиксираните ефекти са $\boldsymbol{x_{ij}}$ и векторите с предсказващите променливи за случайните ефекти са $\boldsymbol{z_{ij}}, i = 1, \ldots, n, j = 1, \ldots, p$.

Нека с вектора $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{ip})', i = 1, \ldots, n$ означим набора от всички ненаблюдавани скрити променливи над обекта i, с вектора $\boldsymbol{b} = (\boldsymbol{b_1}, \boldsymbol{b_2}, \ldots, \boldsymbol{b_n})'$ всички случайни ефекти и с вектора $\boldsymbol{y} = (\boldsymbol{y_1'}, \boldsymbol{y_2'}, \ldots, \boldsymbol{y_n'})'$ набора от всички ненаблюдавани скрити величини над всички обекти.

Оценяването на всички параметри на модела и всички прагове води до неяснозначност. По тази причина се спираме на следните идентификационни ограничения: първите прагове $\alpha_{j,1}, j = 1, \ldots, p$ фиксираме в нулата и дисперсията на нормалните грешки σ^2 фиксираме в 1.

3.2 Намиране на МПО чрез EM алгоритъм

Дефинираме разликите между съседните прагове с $\delta_{j,i} = \alpha_{j,i} - \alpha_{j,i-1}, \ i = 2, \ldots, m_j - 1, j = 1, \ldots, p$. Оттук следва връзката, че $\alpha_{j,i} = \sum_{k=2}^{i} \delta_{j,k}, \ j = 1, \ldots, p, \ i = 2, \ldots, m_j - 1$. Полагаме $\delta_{j,1} = \delta_{j,m_j} = 1$ и $\alpha_{j,0} = 0, j = 1, \ldots, p$. Следващата стъпка е да разгледаме нови променливи, които са линейна трансформация на скритите величини: $y_{ij_{new}} = (y_{ij} - \alpha_{j,y_{ij}^*-1})/\delta_{j,y_{ij}^*}, j = 1, \ldots, p$ и нека $\boldsymbol{y_{i_{new}}} = (y_{i1_{new}}, y_{i2_{new}}, \ldots, y_{ip_{new}})'$. Понеже новите променливи са линейни трансформации на нормални сл. вел., то те също имат нормално разпределе-

ние. Променливата $\boldsymbol{y_{i_{new}}}$, условно по наблюдаваните категорни променливи $\boldsymbol{y^*_{i_{new}}}$, има $p-$мерно орязано нормално разпределение.

3.2.1 Лог-правдоподобие на пълния набор от данни

Пълното лог-правдоподобие $\ln L$ е :

$$\ln L = \ln f(\boldsymbol{b}, \boldsymbol{y_{new}}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{i_{new}}}|\boldsymbol{b_i}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) \prod_{j=1}^{p} f(y_{ij_{new}}|\boldsymbol{b_i}),$$

където $\boldsymbol{y_{new}} = (\boldsymbol{y'_{1_{new}}}, \boldsymbol{y'_{2_{new}}}, \ldots, \boldsymbol{y'_{n_{new}}})'$.

Като изключим константите, пълното лог-правдоподобие има следната аналитична форма:

$$\begin{aligned}
\ln L \;=\; & -0.5 \sum_{i=1}^{n} \ln |\boldsymbol{\Sigma}| - 0.5 \sum_{i=1}^{n} \boldsymbol{b'_i} \boldsymbol{\Sigma^{-1}} \boldsymbol{b_i} + \\
& + \sum_{i=1}^{n} \ln \delta_{1,y^*_{i1}} - 0.5 \sum_{i=1}^{n} [\delta_{1,y^*_{i1}} y_{i1_{new}} - (\boldsymbol{x'_{i1}} \boldsymbol{\beta_1} + \boldsymbol{z'_{i1}} \boldsymbol{b_{i1}} - \alpha_{1,y^*_{i1}-1})]^2 \\
& + \ldots \\
& + \sum_{i=1}^{n} \ln \delta_{p,y^*_{ip}} - 0.5 \sum_{i=1}^{n} [\delta_{p,y^*_{ip}} y_{ip_{new}} - (\boldsymbol{x'_{ip}} \boldsymbol{\beta_p} + \boldsymbol{z'_{ip}} \boldsymbol{b_{ip}} - \alpha_{p,y^*_{ip}-1})]^2.
\end{aligned}$$

Диференцирайки лог-правдободобието и приравнявайки първите производни на нула, ние получаваме затворени форми за оценките на неизвестните параметри $\boldsymbol{\Gamma} = (\boldsymbol{\beta'_1}, \boldsymbol{\beta'_2}, \ldots, \boldsymbol{\beta'_p}, \boldsymbol{\Sigma}, \boldsymbol{\delta'_1}, \boldsymbol{\delta'_2}, \ldots, \boldsymbol{\delta'_p})$, където $\boldsymbol{\delta_j} = (\delta_{j,2}, \ldots, \delta_{j,m_j-1})$, $j = 1, \ldots, p$.

3.2.2 Явен вид на МПО

Оценката за ковариационната матрица на случайните ефекти $\boldsymbol{\Sigma}$ е:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=0}^{n} \boldsymbol{b_i} \boldsymbol{b'_i}.$$

Регресионните параметри за фиксираните ефекти $\boldsymbol{\beta_j}, j = 1, \ldots, p$ удовлетворяват следната система от уравнения:

$$\sum_{i=1}^{n} \boldsymbol{x_{ij}} \boldsymbol{x'_{ij}} \boldsymbol{\beta_j} = \sum_{i=1}^{n} [\delta_{j,y^*_{ij}} y_{ij_{new}} - \boldsymbol{z'_{ij}} \boldsymbol{b_{ij}} + \alpha_{j,y^*_{ij}-1}] \boldsymbol{x_{ij}}.$$

Уравненията за $\delta_{j,k},\ k = 2, \ldots, m_j - 1,\ j = 1, \ldots, p$ са квадратни от вида: $a_j \delta^2_{j,k} + b_j \delta_{j,k} + c_j = 0$, които винаги имат реални корени и по-големият от тях е положителен. Константите $a_j, b_j, c_j, j = 1, \ldots, p$ са:

$$\begin{aligned}
a_j &= \sum_i \sum_{y^*_{ij}=k} (y^2_{ij_{new}}) + n_{j,k+1} + \ldots + n_{j,m}, \\
b_j &= -\sum_i \sum_{y^*_{ij}=k} y_{ij_{new}} (\boldsymbol{x'_{ij}} \boldsymbol{\beta_j} + \boldsymbol{z'_{ij}} \boldsymbol{b_{ij}} - \alpha_{j,k-1}) + \\
&\quad \sum_i \sum_{y^*_{ij}>k} (\delta_{j,y^*_{ij}} y_{ij_{new}} - \boldsymbol{x'_{ij}} \boldsymbol{\beta_j} - \boldsymbol{z'_{ij}} \boldsymbol{b_{ij}} + \delta_{j,2} + \ldots + \delta_{j,k-1} + \delta_{j,k+1} + \ldots + \delta_{j,y^*_{ij}-1}), \\
c_j &= -n_{j,k},
\end{aligned}$$

където $n_{j,k}$ е броят на наблюденията на $j-$тата категорна променлива на $k-$то ниво.

На всяка стъпка на алгоритъма се обновяват оценките на неизвестните параметри. Изчисляването на новите оценки зависи от очакванията на сл.вел., участващи по-горе, при условие наблюдаваните данни. Тези условни математически очаквания зависят само от първите два момента на $p-$мерно орязано нормално разпределение.

3.2.3 Условни очаквания

В дисертацията е показано, че условното разпределение на $\boldsymbol{b_i}$ при условие $\boldsymbol{y_{i_{new}}}$е нормално. В изразите на оценките на E-стъпката на алгоритъма участват следните условни очаквания: $E(\boldsymbol{b_i}|\boldsymbol{y^*_i}), E(\boldsymbol{b_i}\boldsymbol{b'_i}|\boldsymbol{y^*_i}), E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y^*_i})$. Показва се, че те зависят само от първите два момента на $\boldsymbol{y_{i_{new}}}|\boldsymbol{y^*_i}$, което е $p-$мерно орязано нормално разпределение.

3.2.4 $(k+1)$-ва итерация на ECM алгоритъма

Оценките на неизвестните параметри на $k+1-$вата стъпка от предложения ECM алгоритъм са следните:

- $(k+1)$-вата оценка на регресионните параметри $\boldsymbol{\beta_j^{k+1}}, j = 1, \ldots, p$ са решение по МНМК при регресия на $E(\tilde{y}_{ij}|\boldsymbol{y_i^*; \Gamma^k}),\ i = 1, \ldots, n$ по $\boldsymbol{x_{ij}}$.
- $(k+1)$-вата оценка за $\delta_{j,u},\ u = 2, \ldots, m_j - 1, j = 1, \ldots, p$ е: $\delta_{j,u}^{k+1} = (-E[b_j|\boldsymbol{y^*; \Gamma^k}] + \sqrt{(E[b_j|\boldsymbol{y^*; \Gamma^k}]^2 - 4E[a_j|\boldsymbol{y^*; \Gamma^k}]E[c_j|\boldsymbol{y^*; \Gamma^k}])})/2E[a_j|\boldsymbol{y^*; \Gamma^k}]$.

  В изразите за очакванията на a_j, b_j, c_j се използват обновените оценки $\boldsymbol{\beta_j^{k+1}}, j = 1, \ldots, p$ и $\delta_{j,i}^{k+1}, i = 2, \ldots, u-1$.
- $(k+1)$-вата оценка на ковариационната матрица на случайните ефекти е $\hat{\boldsymbol{\Sigma}}^{k+1} = \frac{1}{n}\sum_{i=0}^{n} E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*; \Gamma^k})$, където за изчисленията на очакванията използваме обновените оценки $\boldsymbol{\beta_j^{k+1}}, j = 1, \ldots, p$ и $\delta_{j,i}^{k+1}, i = 2, \ldots, m_j - 1$.

Спиращият критерий е разликата в стойностите на параметрите в две последователни итерации да е по-малка от предварително зададено число.

3.2.5 Приближение на стандартните грешки

Ние използваме "bootstrap" метод, описан в McLachlan and Krishnan [2008] стр. 130-131, за да намерим приближение за стандартните грешки на оценките.

3.3 Симулации

За реализирането на алгоритъма сме използвали безплатната софтуерна среда за статистически изчисления и графики **R**. Условните очаквания, необходими за E-стъпката на алгоритъма, е възможно да бъдат изчислени с аналитични форми според работата на Manjunath and Wilhelm [2009]. При реализацията на алгоритъма е използван този подход, тъй като е достатъчно бърз и осигурява детерминираност на всяка стъпка на алгоритъма за разлика от Монте Карло версията на алгоритъма в предишната глава.

Симулирали сме стойности от следния модел със случайни свободни членове:

$$\begin{aligned} y_{i1} &= \beta_{10} + \beta_{11}x_{i1} + b_{i1} + \epsilon_{i1}, \\ y_{i2} &= \beta_{20} + \beta_{21}x_{i2} + b_{i2} + \epsilon_{i2}, \end{aligned} \tag{3.3}$$

където $\beta_{10} = -0.5, \beta_{11} = 1, \beta_{20} = 1, \beta_{21} = -0.5, Var(\epsilon_{ij}) = 1, j = 1, 2$ с прагове $\alpha_{1,1} = \alpha_{2,1} = 0, \alpha_{1,2} = 1.2, \alpha_{2,1} = 0.7$ и ковариационната матрица на случайните ефекти е:

$$Var\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}.$$

Направихме две симулационни изследвания с различен брой на обектите в извадките ($n = 100$ и $n = 500$). Симулирахме по 100 извадки за всяко изследване. За всяко приближение на стандартните грешки използвахме 50 "bootstrap" извадки. Резултатите са представени в Таблица 3.1.

За втората симулация средните на оценките на параметрите се отличават от истинските стойности на параметрите в рамките на < 0.01. За първото симулационно изследване получаваме изместени оценки, което може да бъде обяснено със свойствата на МПО, които са само асимптотично неизместени.

Както се очаква стандартните грешки намаляват с увеличаване на големината на извадката.

От наблюдението, че стандартните отклонения на оценките и "bootstrap" стандартните грешки са много близки по стойност, можем да заключим, че алгоритъмът схожда, както се очаква.

3.4 Приложение на модела

В периода между началото на 2006 година и началото на 2008 година 121 жени с рак на шийката на матката или рак на тялото на матката са лекувани в Медицински Университет, София. Незаболелите тъкани в областта на тумора са увредени в резултат на лъчелечение. Това води до странични

Таблица 3.1: Оценки и стандартни грешки от двете симулационни изследвания за "probit" модела 3.3 за две наредени категорни величини

параметри	β_{10}	β_{11}	β_{20}	β_{21}	$\delta_{1,2}$	$\delta_{2,2}$	σ_{11}	σ_{12}	σ_{22}
стойности	-0.5	1	1	-0.5	1.2	0.7	1	-0.8	1
Симулация 1: брой обекти $n = 100$									
средно на оценките	$-.550$	1.057	1.070	-0.537	1.233	0.765	1.140	-0.940	1.134
стд. откл. на оценките	0.354	0.316	0.299	0.246	0.217	0.151	0.400	0.428	0.405
средно на "bootstrap" стд. гр.	0.371	0.335	0.291	0.237	0.216	0.167	0.452	0.485	0.456
Симулация 2: брой обекти $n = 500$									
средно на оценките	$-.494$	0.992	1.004	-0.505	1.203	0.703	1.003	-0.802	1.003
стд. откл. на оценките	0.149	0.141	0.116	0.067	0.097	0.067	0.166	0.181	0.170
средно на "bootstrap" стд. гр.	0.166	0.148	0.118	0.084	0.087	0.068	0.160	0.173	0.161

реакции, които са категоризирани в следните видове: кожни, урогенитални и гастроинтестинални. Проучването (което за краткост ще наричаме Раково и Генетично Проучване, РГП) има за цел да изследва връзката между степента на страничните реакции и генетичните характеристики на пациентите. Тези данни са анализирани в Grigorova [2009], където степента на всеки тип реакции е моделирана поотделно с логистична регресия.

Ние разгледахме съвместен "probit" модел за кожните и урогениталните реакции. Променливите взимат стойности от следните нива: отсъстващи реакции (1), слаби реакции (2), умерени или силни реакции (3). Изследвахме как генотипа на полиморфизъм XRCC3 кодон 241 (C>T) е свързан със степента на странични реакции. Променливата XRCC3 241 взима стойност 0 за генотип {C,C} (45 наблюдения) и 1 за генотип {C,T} или {T,T} (76 наблюдения).

Приложихме следния съвместен модел за кожни и урогенитални реакции

от РГП:

$$\begin{aligned} y_{ij} &= \beta_{j0} + \beta_{j1}\text{XRCC3 } 241_i + b_{ij} + \epsilon_{ij}, \\ \text{Реакции}_{ij} &= \begin{cases} \text{Отсъстващи}, y_{ij} \leq \alpha_{j,1}, (\alpha_{j,1} = 0), \\ \text{Слаби}, \ 0 < y_{ij} \leq \alpha_{j,2}, \\ \text{Умерени или силни}, \ y_{ij} > \alpha_{j,2}, \end{cases} \end{aligned} \tag{3.4}$$

където $(\epsilon_{i1}, \epsilon_{i2})' \sim N(\mathbf{0}, \boldsymbol{I_2})$, $j = 1$ за кожни реакции, $j = 2$ за урогенитални и

$$\boldsymbol{\Sigma} = Var \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Таблица 3.2: Таблица с оценки и стандартни грешки от съвместен модел (3.4) и отделни модели за кожни и урогенитални реакции от раково и генетично проучване

	β_{10}	β_{11}	β_{20}	β_{21}	$\delta_{1,2}$	$\delta_{2,2}$	σ_{11}	σ_{12}	σ_{22}
Съвместен модел за кожни и урогенитални реакции 3.4									
оценки	0.880	-0.772	0.554	0.025	1.417	1.500	1.237	0.801	1.366
стд. гр.	0.263	0.264	0.242	0.268	0.187	0.181	0.235	0.307	0.256
z-score	3.342	-2.928	2.292	0.093	7.578	8.275	5.255	2.610	5.327
Отделен модел за кожни реакции									
оценки	0.596	-0.522			0.946				
стд. гр.	0.179	0.212			0.126				
z-score	3.326	-2.457			7.497				
Отделен модел за урогенитални реакции									
оценки			0.362	0.013		0.975			
стд. гр.			0.176	0.210		0.124			
z-score			2.056	0.066		7.822			

Също така оценихме отделни "probit" модели за кожните и урогенитални-те реакции. Резултатите от трите модела са представени в Таблица с оценки и стандартни грешки от съвместен модел (3.4) и отделни модели за кожни и урогенитални реакции от раково и генетично проучване. Таблица 3.2 показ-ва, че всички параметри в съвместния модел 3.4 са статистически значимо отличими от нула освен индикаторната променлива за генотипа на полимор-

физъм XRCC3 кодон 241 (C>T) в подмодела за урогениталните реакции β_{21}.

Таблица 3.3: Таблица със съвместните вероятности за степента на кожните и урогениталните реакции според модела 3.4 в зависимост от генотипа на XRCC3 241

Генотип {C,C}	Урогенитални реакции			
Кожни реакции	отсъстващи	слаби	умерени или силни	маргинални
отсъстващи	0.146	0.094	0.039	0.278
слаби	0.132	0.141	0.089	0.362
умерени или силни	0.081	0.137	0.142	0.360
маргинални	0.359	0.371	0.269	1
Генотип {C,T} или {T,T}	Урогенитални реакции			
Кожни реакции	отсъстващи	слаби	умерени или силни	маргинални
отсъстващи	0.219	0.169	0.083	0.471
слаби	0.100	0.134	0.105	0.338
умерени или силни	0.034	0.069	0.087	0.191
маргинални	0.353	0.372	0.275	1

Параметрите от най-голям интерес са коефициентите β_{11} и β_{21}. Оценката за β_{11} е отрицателна и тест-статистиката на хипотезата, че коефициентът е нула, е $z = -2.928$, p-value= 0.0034. Ефектът на генотипа на полиморфизъм XRCC3 241 върху кожните реакции според данните е статистически значим, докато ефектът върху урогениталните реакции е статистически незначим. Данните ни представят достатъчно доказателства да заключим, че генотип {C,C} на полиморфизъм XRCC3 241 повишава риска от по-тежки кожни реакции.

Вероятностите за силата на кожните и урогениталните реакции според модела 3.4 в зависимост от генотипа на пациентите са представени в Таблица 3.3. Почти равно вероятно е жените с генотип {C,C} на XRCC3 241 да имат 'отсъстващи' кожни и 'отсъстващи' урогенитални реакции, 'слаби' кожни и 'слаби' урогенитални реакции или 'умерени или силни' кожни и 'умерени или силни' урогенитални реакции. Докато при жените с другите два генотипа най-вероятно е да нямат кожни и урогенитални реакции или да имат само 'слаби' урогенитални реакции. Притежаването на алел T има

Таблица 3.4: Таблица с вероятностите за степента на кожните и урогениталните реакции при отделно моделиране в зависимост от генотипа на XRCC3 241

	Степен на реакции		
Генотип {C,C}	отсъстващи	слаби	умерени или силни
Кожни реакции	0.274	0.363	0.363
Урогенитални реакции	0.359	0.373	0.268
	Степен на реакции		
Генотип {C,T} или {T,T}	отсъстващи	слаби	умерени или силни
Кожни реакции	0.468	0.340	0.192
Урогенитални реакции	0.356	0.373	0.271

потискащ реакциите ефект.

Таблица 3.4 представя вероятностите за силата на кожните и урогениталните реакции при отделно моделиране на двата типа реакции. Ако по тези модели се опитаме да направим едновременно извод за степента на проява на кожни и урогенитални реакции, трябва да допуснем, че двата типа реакции са независими и да умножим съответните вероятности. Например за жени с генотип {C,C} на XRCC3 241 вероятността да имат 'отсъстващи' кожни и 'отсъстващи' урогенитални реакции би била $0.274 * 0.359 = 0.098$, а за жени с генотип {C,T} или {T,T} на XRCC3 241 тази вероятност би била $0.471 * 0.353 = 0.166$. Тези вероятности са различни съответно от стойностите 0.146 и 0.219, представени в Таблица 3.3. Забелязва се, че маргиналните вероятности за кожните и урогениталните странични реакции на жените от съвместния модел, независимо от генотипа на полиморфизъм XRCC3 241, са почти същите като тези от отделните модели за двата типа реакции.

Също така при отделните модели не бихме могли да тестваме едновременната значимост на генотипа на полиморфизъм XRCC3 241 за двата типа реакции, без да направим някаква корекция в нивото на значимост за двата теста, докато за такъв тест при модел 3.4 нямаме такива усложнения.

Също така дисперсиите на случайните свободни членове според модел 3.4 са статистически значимо по-големи от 0 ($\text{z-score}_{\sigma_{11}} = 5.255, \text{z-score}_{\sigma_{22}} = 5.327$). Ковариацията между случайните ефекти е статистически значимо по-голяма от нула ($\text{z-score}_{\sigma_{12}} = 2.610$). От това можем да заключим, че зави-

симостта между кожните и урогениталните реакции е положителна.

Наблюдаваната положителна връзка между двата типа реакции е ново откритие. Възможно обяснение е начинът на клиничното установяване на страничните реакции. Някои странични ефекти, отчетени от пациентите като урогенитални, може да са свързани с кожните реакции в третираната зона. Например дизурията е урогенитален страничен ефект и представлява болезнено и затруднено уриниране. Но когато пациентът има кожни раздразнения около гениталната област, това също може да причини болка при уриниране, така че може да се отчетат урогенитални вместо кожни странични реакции.

3.5 Заключение

В тази глава от дисертацията разгледахме корелиран "probit" модел за анализиране на няколко наредени категорни променливи. Описахме разширение на ECM алгоритъма от предишната глава за намиране на МПО на неизвестните параметри в модела. Алгоритъмът е приложно осъществен в средата за статистическа обработка и анализ на данни **R** (R Core Team [2013]). Изследвахме реализацията чрез симулации. Илюстрирахме подхода върху данни, които са анализирани от Grigorova [2009].

Предложеният подход има предимства пред останалите начини за намиране на МПО заради това, че дава асимптотично неизместени оценки и може да се справи с голяма размерност на многомерния отклик.

Скоростта на сходимост на алгоритъма може да бъде увеличена чрез разширение на множеството на оценяваните параметри (Liu et al. [1998]).

Използваният "bootstrap" метод за намиране на приближение на стандартните грешки е изключително времеемък. Може да бъде разгледан метода за приближение на Луис (Louis [1982]).

Друга насока за развитие на алгоритъма е модел с корелирани грешки, както и моделирането на няколко категорни променливи от дългосрочни проучвания.

Глава 4

Съвместен модел за една наредена категорна и една нормална променливи, проследени неколкократно във времето

В тази глава е разгледан корелиран "probit" модел за една наредена категорна и една нормална променливи от дългосрочни проучвания. За намиране на МПО на неизвестните параметри на модела е предложено разширение на ECM алгоритмите, представени в предходните две глави. Резултатите, изложени тук, не са публикувани до момента.

4.1 Описание на модела

Нека са наблюдавани една наредена категорна променлива с m нива, означена с y_1^* и една нормална величина y_2. Наблюденията над $i-$ти обект в момент j са отбелязвани с $\boldsymbol{y_{ij}^*} = (y_{1ij}^*, y_{2ij})'$. Предполагаме, че има скрита нормална величина y_{1ij}, която генерира наблюдаваната категорна променлива y_{1ij}^*. Разглеждаме следния корелиран "probit" модел със случайни ефекти за скритата

и наблюдаваната нормална променливи:

$$\begin{aligned} y_{1ij} &= \boldsymbol{x'_{1ij}}\boldsymbol{\beta_1} + \boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}} + \epsilon_{1ij}, \\ y_{2ij} &= \boldsymbol{x'_{2ij}}\boldsymbol{\beta_2} + \boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}} + \epsilon_{2ij}. \end{aligned} \tag{4.1}$$

Връзката между скритата нормална и наблюдаваната категорна променлива е следната:

$$y^*_{1ij} = \begin{cases} 1, \ y_{1ij} \leq \alpha_1; \\ l, \ \alpha_{l-1} < y_{1ij} \leq \alpha_l, \ l = 2, \ldots, m-1; \\ m, \ y_{1ij} > \alpha_{m-1}; \end{cases} \tag{4.2}$$

за неизвестни прагове $\alpha_1, \ldots, \alpha_{m-1}$.

Предполагаме, че векторът от случайните ефекти е нормално разпределен с размерност q и го означаваме с $\boldsymbol{b_i} = (\boldsymbol{b_{1i}}', \boldsymbol{b_{2i}}')' \sim N(\boldsymbol{0_q}, \boldsymbol{\Sigma})$, където $\boldsymbol{0_q}$ е $q-$мерен нулев вектор-стълб. Допускаме също, че грешките са нормално разпределени $(\epsilon_{1ij}, \epsilon_{2ij})' \sim N(\boldsymbol{0_2}, \boldsymbol{\Sigma_\epsilon})$, независими за отделните индивиди и независими от случайните ефекти.

Ковариационната матрица на грешките е:

$$\boldsymbol{\Sigma_\epsilon} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

а тази на случайните ефекти с размерност $q \times q$ е:

$$\boldsymbol{\Sigma} = Var \begin{pmatrix} \boldsymbol{b_{1i}} \\ \boldsymbol{b_{2i}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}.$$

Регресионните параметри за фиксираните ефекти в модел 4.1 означаваме с p_1- и p_2-мерните вектори $\boldsymbol{\beta_1}$ и $\boldsymbol{\beta_2}$. Векторите с предсказващите променливи за фиксираните ефекти са съответно $\boldsymbol{x_{1ij}}$ и $\boldsymbol{x_{2ij}}$. Векторите с предсказващите променливи за случайните ефекти са $\boldsymbol{z_{1ij}}$ и $\boldsymbol{z_{2ij}}$.

Нека въведем следните означения: $\boldsymbol{y_{1i}} = (y_{1i1}, y_{1i2}, \ldots, y_{1in_i})'$, $\boldsymbol{y_{2i}} = (y_{2i1}, y_{2i2}, \ldots, y_{2in_i})'$, $i = 1, \ldots, n$, $\boldsymbol{b} = (\boldsymbol{b_1}', \boldsymbol{b_2}', \ldots, \boldsymbol{b_n}')'$.

Ако всички неизвестни параметри и прагове се оставят без ограничения,

това води до неоднозначна определеност на модела. Затова налагаме идентификационни ограничения: първия праг α_1 определяме, че е нула и условната дисперсия $\sigma_{1|2} = \sigma_{11} - \sigma_{12}^2/\sigma_{22}$ полагаме да е 1.

4.2 МПО чрез ЕМ алгоритъм

Дефинираме разликите между съседните прагове с $\delta_i = \alpha_i - \alpha_{i-1},\ i = 2, \ldots, m-1$, откъдето следва, че $\alpha_i = \sum_{k=2}^{i} \delta_k,\ i = 2, \ldots, m-1$. За по-нататъшна употреба определяме $\delta_1 = \delta_m = 1$ и $\alpha_0 = 0$. Следващата стъпка е да разгледаме нова променлива, която е линейна трансформация на скритата величина: $y_{1ij_{new}} = (y_{1ij} - \alpha_{y^*_{1ij}-1})/\delta_{y^*_{1ij}}$. По-конкретно, ако $y^*_{1ij} = u,\ u = 1, \ldots, m$, тогава $y_{1ij_{new}} = (y_{1ij} - \alpha_{u-1})/\delta_u$. Понеже новата променлива е линейна трансформация на нормална сл. вел., то тя също има нормално разпределение.

4.2.1 Лог-правдоподобие на пълния набор от данни

Пълното лог-правдоподобие $\ln L$ е:

$$\begin{aligned} \ln L &= \ln f(\boldsymbol{b}, \boldsymbol{y_{1_{new}}}, \boldsymbol{y_2}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}} | \boldsymbol{b_i}) \\ &= \sum_{i=1}^{n} \ln[f(\boldsymbol{b_i}) \prod_{j=1}^{n_i} f(y_{2ij}|\boldsymbol{b_i}) f(y_{1ij_{new}}|\boldsymbol{b_i}, y_{2ij})], \end{aligned}$$

където $\boldsymbol{y_{1i_{new}}} = (y_{1i1_{new}}, y_{1i2_{new}}, \ldots, y_{1in_{i_{new}}})'$, $\boldsymbol{y_{1_{new}}} = (\boldsymbol{y'_{11_{new}}}, \boldsymbol{y'_{12_{new}}}, \ldots, \boldsymbol{y'_{1n_{new}}})'$ и $\boldsymbol{y_2} = (\boldsymbol{y'_{21}}, \boldsymbol{y'_{22}}, \ldots, \boldsymbol{y'_{2n}})'$.

Като изключим константите, пълното лог-правдоподобие има следната аналитична форма:

$$\begin{aligned} \ln L &= -0.5 \sum_{i=1}^{n} \ln |\boldsymbol{\Sigma}| - 0.5 \sum_{i=1}^{n} \boldsymbol{b'_i} \boldsymbol{\Sigma^{-1}} \boldsymbol{b_i} \\ &\quad -0.5 \sum_{i=1}^{n} \sum_{j=1}^{n_i} \ln \sigma_{22} - 0.5 \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{(y_{2ij} - \mu_{2ij})^2}{\sigma_{22}} \end{aligned}$$

$$-0.5\sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\sigma_{1|2}+\sum_{i=1}^{n}\sum_{j=1}^{n_i}\ln\delta_{y^*_{1ij}}-0.5\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\delta^2_{y^*_{1ij}}(y_{1ij}-\mu_{1ij})^2}{\sigma_{1|2}},$$

където

$$\begin{aligned}\lambda &= \sigma_{12}/\sigma_{22},\\ \mu^*_{1ij} &= (\boldsymbol{x'_{1ij}}\boldsymbol{\beta_1}+\boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}}-\alpha_{y^*_{1ij}-1})/\delta_{y^*_{1ij}},\\ \mu_{2ij} &= \boldsymbol{x'_{2ij}}\boldsymbol{\beta_2}+\boldsymbol{z'_{2ij}}\boldsymbol{b_{2i}},\\ \mu_{1ij} &= \mu^*_{1ij}+\lambda(y_{2ij}-\mu_{2ij})/\delta_{y^*_{1ij}}.\end{aligned}$$

Явен вид на оценките на неизвестните параметри $\boldsymbol{\Gamma}=(\boldsymbol{\beta_1},\boldsymbol{\beta_2},\boldsymbol{\Sigma},\boldsymbol{\delta},\sigma_{22},\lambda)$, където $\boldsymbol{\delta}=(\delta_2,\ldots,\delta_{m-1})$, получаваме като диференцираме функцията на пълното лог-правдоподобие и приравним първите производни на нула.

4.2.2 Явен вид на МПО

Оценката за ковариационната матрица на случайните ефекти $\boldsymbol{\Sigma}$ е:

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{i=0}^{n}\boldsymbol{b_i}\boldsymbol{b'_i}.$$

Оценката за дисперсията σ_{22} на грешката на наблюдаваната нормална величина е:

$$\hat{\sigma}_{22}=\frac{1}{\sum_{i=1}^{n}n_i}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{2ij}-\mu_{2ij})^2.$$

Регресионните параметри за фиксираните ефекти $\boldsymbol{\beta_1}$ за скритата нормална променлива удовлетворяват следната система от уравнения:

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\boldsymbol{x_{1ij}}\boldsymbol{x'_{1ij}}\boldsymbol{\beta_1}=\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{y^*_{1ij}}y_{1ij_{new}}-\boldsymbol{z'_{1ij}}\boldsymbol{b_{1i}}+\alpha_{y^*_{1ij}-1}-\lambda(y_{2ij}-\mu_{2ij})]\boldsymbol{x_{1ij}}.$$

Регресионните параметри за фиксираните ефекти $\boldsymbol{\beta_2}$ за наблюдаваната

нормална променлива удовлетворяват следната система от уравнения:

$$(1+\lambda^2\sigma_{22})\sum_{i=1}^{n}\sum_{j=1}^{n_i}\boldsymbol{x_{2ij}x'_{2ij}\beta_2} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\lambda^2\sigma_{22})(y_{2ij}-\boldsymbol{z'_{2ij}b_{2i}})\boldsymbol{x_{2ij}} - \lambda\sigma_{22}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[\delta_{y^*_{1ij}}y_{1ij_{new}}-\boldsymbol{x'_{1ij}\beta_1}-\boldsymbol{z'_{1ij}b_{1i}}+\alpha_{y^*_{1ij}-1}]\boldsymbol{x_{2ij}}.$$

Уравненията за $\delta_k,\ k=2,\ldots,m-1$ са квадратни уравнения от вида: $a\delta_k^2+b\delta_k+c=0$, които винаги имат реални корени и по-големият от тях е винаги положителен. Константите a,b,c са:

$$\begin{aligned}
a &= \sum_{i,j}\sum_{y^*_{1ij}=k}(y^2_{ij_{new}})+n_{k+1}+\ldots+n_m,\\
b &= -\sum_{i,j}\sum_{y^*_{1ij}=k}y_{1ij_{new}}(\boldsymbol{x'_{1ij}\beta_1}+\boldsymbol{z'_{1ij}b_{1i}}-\alpha_{k-1}+\lambda(y_{2ij}-\mu_{2ij}))+\\
&\quad \sum_{i,j}\sum_{y^*_{1ij}>k}(\delta_{y^*_{1ij}}y_{1ij_{new}}-\boldsymbol{x'_{1ij}\beta_1}-\boldsymbol{z'_{1ij}b_{1i}}+\alpha_{-k,y^*_{1ij}-1}-\lambda(y_{2ij}-\mu_{2ij})),\\
c &= -n_k,
\end{aligned}$$

където n_k е броят на наблюденията на $k-$то ниво на категорната величина и $\alpha_{-k,y^*_{1ij}-1}=\delta_2+\ldots+\delta_{k-1}+\delta_{k+1}+\ldots+\delta_{y^*_{1ij}-1}$.

Оценката на параметъра λ удовлетворява следното уравнение:

$$\lambda\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{2ij}-\mu_{2ij})^2=\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\delta_{y^*_{1ij}}y_{1ij_{new}}-\delta_{y^*_{1ij}}\mu^*_{1ij})(y_{2ij}-\mu_{2ij}).$$

Очакванията на сл.вел., участващи в явния вид на МПО, при условие наблюдаваните данни зависят само от първите два момента на многомерно орязано нормално разпределение.

4.2.3 Условни очаквания

В дисертацията е изведено условното разпределение на $\boldsymbol{b_i}$ при условие $\boldsymbol{(y_{1i_{new}}, y_{2i})}$, което е нормално. В изразите на оценките на Е-стъпката на алгоритъма участват следните условни очаквания: $E(\boldsymbol{b_i}|\boldsymbol{y^*_{1i}, y_{2i}})$, $E(\boldsymbol{b_ib'_i}|\boldsymbol{y^*_{1i}, y_{2i}})$, $E(y_{1ij_{new}}\boldsymbol{b_i}|\boldsymbol{y^*_{1i}, y_{2i}})$.

В дисертацията е показано, че те зависят само от първите два момента на $\boldsymbol{y_{1i_{new}}|(y^*_{1i}, y_{2i})}$, което е орязано нормално разпределение.

4.2.4 $(k+1)$-ва итерация на ECM алгоритъма

Оценките на неизвестните параметри на $k+1-$вата стъпка $\boldsymbol{\Gamma^{k+1}}$ от предложения ECM алгоритъм са следните:

- $(k+1)$-вата оценка на регресионните параметри за скритата нормална величина $\boldsymbol{\beta_1^{k+1}}$ са решение по МНМК при регресия на $E(\tilde{y}_{1ij}|\boldsymbol{y^*_{1i}, y_{2i}; \Gamma^k})$ по $\boldsymbol{x_{1ij}}$.

- $(k+1)$-вата оценка на дисперсията на грешките в подмодела за наблюдаваната непрекъсната величина е:

$$\hat{\sigma}_{22}^{k+1} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{j=1}^{n_i} E[(y_{2ij} - \mu_{2ij})^2 | \boldsymbol{y^*_{1i}, y_{2i}; \Gamma^k}].$$

- $(k+1)$-вата оценка на параметъра λ е:

$$\lambda^{k+1} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} E[(\delta_{y^*_{1ij}} y_{1ij_{new}} - \delta_{y^*_{1ij}} \mu^*_{1ij})(y_{2ij} - \mu_{2ij}) | \boldsymbol{y^*_{1i}, y_{2i}; \Gamma^k}]}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} E[(y_{2ij} - \mu_{2ij})^2 | \boldsymbol{y^*_{1i}, y_{2i}; \Gamma^k}]}.$$

  За изчисляването на λ^{k+1} се използват обновените оценки $\boldsymbol{\beta_1^{k+1}}$.

- $(k+1)$-вата оценка за $\delta_u,\ u = 2, \ldots, m-1$ е: $\delta_u^{k+1} =$

$$\frac{-E[b|\boldsymbol{y^*_1, y_2; \Gamma^k}] + \sqrt{(E[b|\boldsymbol{y^*_1, y_2; \Gamma^k}]^2 - 4E[a|\boldsymbol{y^*_1, y_2; \Gamma^k}]E[c|\boldsymbol{y^*_1, y_2; \Gamma^k}])}}{2E[a|\boldsymbol{y^*_1, y_2; \Gamma^k}]}.$$

  В изразите за очакванията на a, b, c се използват вече обновените оценки $\boldsymbol{\beta_1^{k+1}}$, $\lambda^{k+1}, \hat{\sigma}_{22}^{k+1}, \delta_i^{k+1}, i = 2, \ldots, u-1$.

- $(k+1)$-вата оценка на регресионните параметри за наблюдаваната нормална величина $\boldsymbol{\beta_2^{k+1}}$ са решение на следната система от линейни урав-

нения:

$$(1+\lambda^2\sigma_{22})\sum_{i=1}^{n}\sum_{j=1}^{n_i}\boldsymbol{x_{2ij}x'_{2ij}\beta_2} \quad = \\ \sum_{i=1}^{n}\sum_{j=1}^{n_i}(1+\lambda^2\sigma_{22})E[(y_{2ij}-\boldsymbol{z'_{2ij}b_{2i}})|\boldsymbol{y^*_{1i}, y_{2i}; \Gamma^k}]\boldsymbol{x_{2ij}} \\ -\lambda\sigma_{22}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E[\delta_{y^*_{1ij}}y_{1ij_{new}}-\boldsymbol{x'_{1ij}\beta_1}-\boldsymbol{z'_{1ij}b_{1i}}+\alpha_{y^*_{1ij}-1}|\boldsymbol{y^*_{1i}, y_{2i}; \Gamma^k}]\boldsymbol{x_{2ij}}.$$

За пресмятането на стойността на $\boldsymbol{\beta_2^{k+1}}$ се използват вече обновените оценки $\boldsymbol{\beta_1^{k+1}}$, $\lambda^{k+1}, \hat{\sigma}_{22}^{k+1}, \delta_u^{k+1}, u = 2, \ldots, m-1$.

- $(k+1)$-вата оценка на ковариационната матрица на случайните ефекти е $\hat{\boldsymbol{\Sigma}}^{k+1} = \frac{1}{n}\sum_{i=0}^{n}E(\boldsymbol{b_i b'_i}|\boldsymbol{y^*_{1i}, y_{2i}; \Gamma^k})$. За пресмятането на стойността на $\hat{\boldsymbol{\Sigma}}^{k+1}$ се използват вече обновените оценки $\boldsymbol{\beta_1^{k+1}}, \boldsymbol{\beta_2^{k+1}}, \lambda^{k+1}, \hat{\sigma}_{22}^{k+1}, \delta_u^{k+1}, u = 2, \ldots, m-1$.

Алгоритъмът спира при $|\boldsymbol{\Gamma^{k+1}} - \boldsymbol{\Gamma^k}| < \epsilon$, където ϵ е предварително избрано малко число.

4.2.5 Приближение на стандартните грешки

Ние използваме "bootstrap" метод, описан в McLachlan and Krishnan [2008] стр. 130-131, за да намерим приближение за стандартните грешки на оценките.

4.3 Симулации

Симулирани са стойности от следния модел със случайни свободни членове:

$$\begin{aligned} y_{1ij} &= \beta_{10} + \beta_{11}t_{ij} + b_{1i} + \epsilon_{1ij}, \ j = 1, \ldots, 4, \\ y_{2ij} &= \beta_{20} + \beta_{21}t_{ij} + b_{2i} + \epsilon_{2ij}, \ j = 1, \ldots, 4, \end{aligned} \tag{4.3}$$

където $\beta_{10} = -0.5, \beta_{11} = 1, \beta_{20} = 1, \beta_{21} = -0.5$, с прагове $\alpha_1 = 0, \alpha_2 = 1.2$ за категорна променлива с три нива, $\lambda = 1/(2\sqrt{3}) \approx 0.2887$. Ковариационната матрица на грешките е:

$$\mathbf{\Sigma_\epsilon} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 4/3 & 2/\sqrt{3} \\ 2/\sqrt{3} & 4 \end{pmatrix}.$$

Ковариационната матрица на случайните ефекти е:

$$\mathbf{\Sigma} = Var \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} = \begin{pmatrix} \sigma^b_{11} & \sigma^b_{12} \\ \sigma^b_{21} & \sigma^b_{22} \end{pmatrix} = \begin{pmatrix} 1 & \text{-0.8} \\ \text{-0.8} & 1 \end{pmatrix}.$$

Симулирахме 100 извадки с два различни броя индивиди в извадката ($n = 100$ и $n = 200$) с по 4 наблюдения над обект. За всяко приближение на стандартните грешки сме използвали 50 "bootstrap" извадки. Резултатите са представени в Таблица 4.1.

Таблица 4.1: Таблица с оценките и стандартните грешки от двете симулационни изследвания за "probit" модела 4.3 за повторни наблюдения на една наредена категорна и една нормална величини

пар	β_{10}	β_{11}	β_{20}	β_{21}	δ_1	λ	σ_{22}	σ^b_{11}	σ^b_{12}	σ^b_{22}
ст-ти	-0.5	1	1	-0.5	1.2	0.289	4	1	-0.8	1
Симулация 1: брой обекти $n = 100$										
ср. на оценк.	-0.513	1.025	0.984	-0.498	1.214	0.290	3.888	1.104	-0.778	0.991
стд.от. на оц.	0.225	0.106	0.277	0.093	0.131	0.053	0.275	0.346	0.223	0.297
ср. на стд.гр.	0.235	0.112	0.255	0.085	0.138	0.055	0.296	0.388	0.210	0.267
Симулация 2: брой обекти $n = 200$										
ср. на оценк.	-.525	1.014	1.015	-0.503	1.197	0.288	3.983	1.023	-0.808	1.027
стд.от. на оц.	0.136	0.064	0.186	0.068	0.090	0.033	0.231	0.216	0.142	0.181
ср. на стд.гр.	0.160	0.074	0.187	0.064	0.091	0.036	0.221	0.233	0.141	0.191

С увеличаване на големината на извадката средните на оценките за параметрите се приближават към истинските стойности на параметрите с изключение на свободния член за категорната величина и дисперсията на случайния ефект за непрекъснатата променлива, като отдалечаването е по-малко от 0.02 по абсолютна стойност. Но дисперсията на случайния ефект за скритата величина е много по-близо до истинската стойност (приближението е по-добро с приблизително 0.1), а дисперсията на грешката са непрекъснатата величина е по-близо с 0.08. Поради ограничени ресурси големините на извадките не са големи, като очакванията са при голяма извадка ($n = 1000$) оценките да са неизместени.

Стандартните отклонения за по-голямата извадка са по-малки от стандартните отклонения на по-малката извадка, което е очакван резултат. Приблизителното равенство между средното на "bootstrap" стандартните грешки и стандартните отклонения на оценките е показател, че алгоритъмът работи коректно.

4.4 Приложение на модела

Прилагаме предложения модел 4.1 към данните от ЗПП (използвани в секция Приложение на модела на Глава 2). Променливите от основен интерес са самооценката за здравето на индивидите и индексът на телесна маса. Ще изследваме как двете характеристики се променят във времето. Оценяваме следния модел със случайни свободни членове:

$$\begin{aligned} y_{ij} &= \beta_{10} + \beta_{11} t_{ij} + b_{1i} + \epsilon_{1ij}, j = 1, \ldots, 7 \\ \text{bmi}_{ij} &= \beta_{20} + \beta_{21} t_{ij} + b_{2i} + \epsilon_{2ij}, j = 1, \ldots, 7 \\ \text{srh}_{ij} &= \begin{cases} 1, \ y_{ij} \leq \alpha_1 = 0, \\ j, \ \alpha_{j-1} < y_{ij} \leq \alpha_j, \ j = 2, \ldots, m-1, \\ 5, \ y_{ij} > \alpha_{m-1}, \end{cases} \end{aligned} \tag{4.4}$$

където ковариационната матрица на грешките е:

$$\mathbf{\Sigma_\epsilon} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

а тази на случайните ефекти е:

$$\mathbf{\Sigma} = Var \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} = \begin{pmatrix} \sigma^b_{11} & \sigma^b_{12} \\ \sigma^b_{21} & \sigma^b_{22} \end{pmatrix}.$$

С 'bmi_{ij}' и 'srh_{ij}' сме отбелязали съответно индекса на телесна маса и самооценката за здравето на $i-$ти индивид в момент от време j.

В анализа са включени 7244 индивида от изследването, които имат пълен набор от наблюдения. Резултатите са представени в Таблица 4.2.[1]

Таблица 4.2: Таблица с оценки и стандартни грешки от съвместния корелиран "probit" модел 4.4 за самооценка за здравето и индекса на телесна маса, приложен към данните от здравно и пенсионно проучване.

параметри	β_{10}	β_{11}	β_{20}	β_{21}	δ_2	δ_3
оценки	1.216	0.115	26.920	0.149	1.582	1.494
стандартни грешки	0.0171	0.0026	0.0366	0.0039	0.0107	0.0110
z-score	71.25	44.23	735.12	38.33	148.50	135.39
параметри	δ_4	λ	σ_{22}	σ^b_{11}	σ^b_{12}	σ^b_{22}
оценки	1.414	0.013	3.275	2.158	1.980	23.661
стандартни грешки	0.0180	0.0028	0.0236	0.0400	0.0908	0.4173
z-score	78.63	4.83	139.03	53.97	21.80	56.70

Таблица 4.2 показва, че всички параметри в модела са статистически значимо отличими от нула, сред които и параметрите, които описват зависимост между самооценката за здравето и индекса на телесна маса (λ и σ^b_{12}). От най-голям интерес са регресионните коефициенти β_{11} и β_{21}. И двата коефициента са положителни и свидетелстват, че с времето самооценката за здравословното състояние на индивидите се занижава и индексът на телесна маса се увеличава. Дисперсиите на случайните свободни членове са статистически

[1] Оценките са получени с Монте Карло версия на алгоритъма. Стандартните грешки са оценени с 60 "bootstrap" извадки.

значимо различни от 0, а корелацията между тях е положителна. Това предполага, че има различия между обектите в самооценките за здравето и индекса на телесна маса, както и че има зависимост между наблюденията над един индивид.

Предимството на съвместното моделиране е, че можем едновременно да предскажем двете променливи от интерес, докато при отделно моделиране това е възможно, само ако допуснем, че няма зависимост между двете величини. Такова допускане не винаги е възможно. Друго предимство на съвместния модел е възможността за едновременно тестване на няколко параметъра, докато при тестване на параметри от отделни модели се налага корекция на нивото на значимост за всеки тест.

4.5 Заключение

В тази глава от дисертацията е разгледан корелиран "probit" модел за анализиране на една наредена категорна променлива и една нормална величина от дългосрочни проучвания. Описано е разширение на ECM алгоритмите от главите Модел за една наредена категорна променлива с повторни наблюдения и Съвместен модел за няколко наредени категорни променливи за намиране на МПО на неизвестните параметри в модела. Алгоритъмът е приложно осъществен в средата за статистическа обработка и анализ на данни **R** (R Core Team [2013]). Представени са резултатите от две симулационни изследвания, които потвърждават надеждността на представения метод за оценяване на модела. Подходът е илюстриран върху данни от ЗПП.

Предложеният подход има предимства пред останалите начини за намиране на МПО заради това, че дава асимптотично неизместени оценки и изчисленията не нарастват експоненциално с нарастване на размерността на случайните ефекти.

Скоростта на сходимост на алгоритъма може да бъде увеличена чрез разширение на множеството на оценяваните параметри (Liu et al. [1998]).

Използваният "bootstrap" метод за намиране на приближение на стандартните грешки е изключително времеемък. Може да бъде разгледан методът за приближение на Луис (Louis [1982]).

Глава 5

Заключение

5.1 Научен и практически принос на дисертацията

В дисертацията са изведени ECM алгоритми за намиране на МПО на неизвестните параметри на три различни корелирани "probit" модела. Първият модел е за наредена категорна променлива от дългосрочно проучване. Вторият модел е за няколко наредени категорни променливи. Накрая е разгледан модел за две величини с повторни наблюдения. Едната е наредена категорна, а другата е нормална.

Приложен принос на дисертацията е практическото осъществяване на предложените алгоритми в безплатната среда за статистическа обработка и анализ на данни **R**, както и използването им за оценяване на модели за реални данни.

5.2 Бъдещи насоки за развитие

При продължителни във времето проучвания част от обектите отпадат в различни етапи по различни причини. Анализирайки само наличните данни, оценките на параметрите в модела могат да бъдат изместени. Моделите със случайни ефекти дават неизместени оценки само в случаите, когато механизмът на липсите на данни е неинформативен (т.е. независещ от данните,

които не сме наблюдавали). Когато данните липсват въз основа на някакъв систематичен признак, се налага допълнителен анализ, който взима предвид механизма на липсващите данни.

За решаване на проблема с изместването на оценките при моделиране на една променлива, проследявана неколкократно във времето, са предложени съвместни модели на тази променлива и времето до отпадане на обекта от изследването. Статиите на Diggle et al. [2008]; Henderson et al. [2000]; Tsiatis and Davidian [2004] разглеждат различни методи за съвместното моделиране на една, мерена във времето, променлива и времето до настъпване на събитие. При тези модели най-честият модел за времето до отпадане е полу-параметричният модел на Кокс с пропорционални рискове (Therneau and Grambsch [2001]). Вариациите на моделите на Кокс имат интуитивна интерпретация и се оценяват лесно, когато времената до настъпване на съответното събитие са точно наблюдавани. Когато е известен само интервал от време, в който е настъпило събитието, параметричните модели се считат за по-подходящи. Напълно параметрични модели са разгледани в статията на Sparling et al. [2006]. Предложените модели обхващат модела на Вайбул, отрицателно биномния и лог-логистичния модел като частни случаи.

Като бъдещи насоки на развитие може да се разгледа съвместен модел на две променливи от дългосрочни проучвания заедно с моделиране на времето до отпадане на обектите от изследването, за да се избегне изместване на оценките на параметрите, описващи променливите, менящи се във времето. Моделът за оцеляване може да е предложеният от Sparling et al. [2006] и разгледан от Gueorguieva et al. [2012].

Публикувани и докладвани резултати, изложени в дисертационния труд

Доклади върху резултати от дисертационния труд:

[1] Grigorova, D., *Assessing the effect of genetic factors and other factors on the normal tissue reactions after radiotherapy in patients with cancer*, 16th European Young Statisticians Meeting, Romania, Bucharest, 24-28 August 2009, (Abstract on page 24 in the Book of Abstracts);

[2] Grigorova, D., *Implementation of EM algorithm for maximum likelihood estimation of joint model of one ordinal and one continuous outcome*, Bulgaria, Pomorie, 23-30 June 2012, (Abstract on page 17 in the Book of Abstracts);

[3] Grigorova, D., *Correlated probit model for multiple side effects in cancer radiotherapy*, Annual International Conference on Mathematical Methods and Models in Biosciences, Bulgaria, Sofia, 16-21 June 2013, (BIOMATH 2013, http://www.biomath.bg/2013/index.php), (Abstract on page 51 in the Book of Abstracts);

[4] Grigorova, D., *Correlated probit model for estimation of the relationship between genotypes and multiple side effects in cancer radiotherapy*, 34th Annual Conference of the International Society for Clinical Biostatistics, Germany, Munich, 25-29 August 2013, *Conference Award for Scientist*, (http://www.iscb2013.info/), (Abstract C31.6 in the Book of Abstracts);

[5] Григорова, Д., *ЕМ алгоритми за МПО на корелирани „probit“ модели със случайни ефекти*, Първа докторантска конференцията по математика, информатика и обучение, България, София, 19-20 Септември 2013

(http://mie.uni-sofia.bg/);

[6] Григорова, Д., *ЕМ алгоритми за МПО на корелирани „probit“ модели със случайни ефекти*, Национален семинар по теория на вероятностите и математическа статистика, България, София, 30 Октомври 2013 г.

Статии върху резултати от дисертационния труд:

[1] Grigorova, D. and Gueorguieva, R., *Implementation of the EM algorithm for maximum likelihood estimation of a random effects model for one longitudinal ordinal outcome*, Pliska Stud. Math. Bulgar., 22:41-56, 2013;

[2] Grigorova, D. and Gueorguieva, R., *EM Algorithms for MLE of Correlated Probit Models*, In [MIE 2013] Proceedings, pages 17-24, 2013;

[3] Grigorova, D., Encheva, E. and Gueorguieva, R., *EM algorithm for MLE of a probit model for multiple ordinal outcomes*, accepted for publication in Serdica Journal of Computing in December 2013.

Библиография

J. R. Ashford and R. R. Sowden. Multi-variate probit analysis. *Biometrics*, 26 (3):pp. 535–546, 1970. ISSN 0006341X. URL http://www.jstor.org/stable/2529107. 1

C. I. Bliss. The method of probits. *Science*, pages 38–39, 1934a. 1

C. I. Bliss. The method of probits - a correction. *Science*, pages 409–410, 1934b. 1

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993. 3

G. Casella and E. I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992. ISSN 00031305. doi: 10.2307/2685208. URL http://dx.doi.org/10.2307/2685208. 10

P. J. Catalano. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, 16(8):883–900, 1997. ISSN 0277-6715. URL http://www.biomedsearch.com/nih/Bivariate-modelling-clustered-continuous-ordered/9160486.html. 2

J. S. K. Chan and A. Y. C. Kuk. Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics*, 53:86–97, 1997. 7, 14

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):1–22, 1977. 3

P. Diggle, I. Sousa, and A. Chetwynd. Joint Modelling of Repeated Measurements and Time-to-event Outcomes: The Fourth Armitage Lecture. *Statistics in Medicine*, 27:2981–2998, 2008. 38

F. Drasgow. *Polychoric and Polyserial Correlations*, pages 69–74. John Wiley Sons, Inc., 2004. ISBN 9780471667193. doi: 10.1002/0471667196.ess2014.pub2. URL http://dx.doi.org/10.1002/0471667196.ess2014.pub2. 14

D. Dunson, B. Chen, and J. Harry. A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*, 59:521–530, 2003. 3

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics & Applied Probability*. Chapman & Hall, New York, 1 edition, 1994. ISBN 0412042312. URL http://www.worldcat.org/isbn/0412042312. 12

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, second edition, 2001. 1, 2, 3

G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley, 2004. ISBN 9780471214878. URL http://books.google.bg/books?id=gCoTIFejMgYC. 1

J. H. Gaddum. Methods of biological assay depending on a quantal response. *Reports on biological standards. III.*, 1933. 1

H. Geys, M.M. Regan, P.J. Catalano, and G. Molenberghs. Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data. *J. Agric. Biol. Environ. Stat*, 6:340–355, 2001. 3

R. D. Gibbons and D. Hedeker. Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology*, 62(2):285–296, 1994. doi: 10.1037/0022-006X.62.2.285. 2

D. Grigorova. Assessing the effect of genetic factors and other factors on the normal tissue reactions after radiotherapy in patients with cancer. In *Proceedings of the 16th EYSM*, pages 108–112, 2009. Short paper. 21, 25

D. Grigorova and R. Gueorguieva. Implementation of the EM algorithm for maximum likelihood estimation of a random effects model for one longitudinal ordinal outcome. *Pliska Stud. Math. Bulgar.*, 22:41–56, 2013a. 5

D. Grigorova and R. Gueorguieva. EM Algorithms for MLE of Correlated Probit Models. In *[MIE 2013] Proceedings*, pages 17–24, 2013b. 15

L. Grilli and C. Rampichini. Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics*, 28:31–44, 2003. 2, 3

R. Gueorguieva, R. Rosenheck, and H. Lin. Joint modelling of longitudinal outcome and interval-censored competing risk dropout in a schizophrenia clinical trial. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(2):417–433, 2012. ISSN 1467-985X. doi: 10.1111/j.1467-985X.2011.00719.x. URL http://dx.doi.org/10.1111/j.1467-985X.2011.00719.x. 38

R. V. Gueorguieva. Correlated probit model. In *Encyclopedia of Biopharmaceutical Statistics*, chapter 59, pages 355–362. 2006. doi: 10.3109/9781439822463.057. URL http://informahealthcare.com/doi/abs/10.3109/9781439822463.057. 2

R. V. Gueorguieva and A. Agresti. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96:1102–1112, 2001. URL http://EconPapers.repec.org/RePEc:bes:jnlasa:v:96:y:2001:m:september:p:1102-1112. 3

R. V. Gueorguieva and G. Sanacora. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, 25:1307–1322, 2006. 2, 3

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480, 2000. 38

H. Kawakatsu and A. G. Largey. EM algorithms for ordered probit models with endogenous regressors. *Econometrics Journal*, 12:164–186, 2009. 2, 3, 7

C. Liu, D. Rubin, and Y. Wu. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85(4):755–770, 1998. 25, 36

Q. Liu and D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81(3):624–629, 1994. doi: 10.1093/biomet/81.3.624. URL http://biomet.oxfordjournals.org/content/81/3/624.abstract. 3

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 44(2):226–233, 1982. 25, 36

B. G. Manjunath and S. Wilhelm. Moments calculation for the double truncated multivariate normal density. *http://ssrn.com/abstract=1472153*, September 11 2009. URL http://ssrn.com/abstract=1472153. 10, 19

P. McCullagh and J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989. 2

C. McCulloch. Joint modelling of mixed outcome types using latent variables. *Stat Methods Med Res*, 17(1):53–73, 2008. 3

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, March 2008. ISBN 0471201707. 11, 19, 32

X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993. doi: 10.1093/biomet/80.2.267. URL http://biomet.oxfordjournals.org/content/80/2/267.abstract. 10

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0. i, 4, 14, 25, 36

P. A. Ruud. Extensions of estimation methods using the EM algorithm. *Journal of Econometrics*, 49(3):305–341, September 1991. 7

Y. H. Sparling, N. Younes, Lachi. J. M., and O. M. Bautista. Parametric Survival Models for Interval-censored Data with Time-dependent Covariates. *Biostatistics*, 7:599–614, 2006. 38

T. M. Therneau and P. M. Grambsch. *Modeling survival data Extending the Cox model*. Springer-Verlag, New York, 2001. 38

A. A. Tsiatis and M. Davidian. Joint Modelling of Longitudinal and Time to Event Data: An Overview. *Statistica Sinica*, 14:809–834, 2004. 38

R. Wolfinger and M. O'Connell. Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48: 233–243, 1993. 3

Декларация за оригиналност на резултатите

Декларирам, че дисертацията на тема „EM алгоритми за “probit” модели със случайни ефекти“ съдържа оригинални резултати, получени при проведени от мен научни изследвания (с подкрепата и съдействието на научния ми ръководител). Резултатите, които са получени, описани и/или публикувани от други учени, са надлежно и подробно цитирани в библиографията.

Дисертацията не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

16.12.2013 г., Подпис:

София Деница Григорова