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

Деница Григорова

Катедра ВОИС
Факултет по Математика и Информатика
Софийски Университет "Св. Климент Охридски"

Дисертация за придобиване на образователна и научна степен

Доктор

Декември 2013

Научен ръководител:
доц. д-р Ралица Георгиева

© Деница Григорова 2013

Всички права запазени

Посвещавам този труд на хората, които са ме подкрепяли в трудните моменти

Благодарности

На първо място ще изкажа думи на благодарност към научния ми ръководител доц. Ралица Георгиева. Освен на статистика, тя ме научи на търпение, самомотивация, вяра в собствените възможности. Благодарности към доц. Марусия Божкова за това, че беше първият човек, който ми гласува доверие и ми даде шанс да се развивам в академичните среди. Благодарности към доц. Пламен Матеев, ас. Нина Даскалова, доц. Весела Стоименова, ас. Емил Каменов за доброто отношение и подкрепата. Благодарности към състава на цялата катедра ВОИС и членовете на Факултета по Математика и Информатика за това, че ме приеха като част от тях. Благодаря и на студентите, с които имах възможност да се срещна през годините и които много подпомогнаха моето развитие в областта на статистиката.

Искам да изкажа дълбока признателност към родителите си. Без тяхната подкрепа, усилия и отдаденост нямаше да постигна тези резултати. Благодаря на близките и приятелите, които са важна част от процеса на моето развитие и са били до мен през дългото време на осъществяването на този труд.

Резюме

В настоящата дисертация са разработени EM 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 EM алгоритъм е за съвместен модел за няколко наредени категорни променливи. За всички EM 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

Съдържание iv

Списък на фигурите vii

Списък на таблиците viii

Означения и съкращения x

1 Въведение 1

1.1 Литературен обзор 2

1.2 Структура на дисертацията 6

2 Модели със случайни ефекти за данни с повторни наблюдения 8

2.1 Линейни модели 9

2.1.1 Линейни модели с фиксирани ефекти 9

2.1.2 Линейни модели със смесени ефекти 9

2.2 Обобщени линейни модели 12

2.2.1 Обобщени линейни модели с фиксирани ефекти 12

2.2.2 Обобщени линейни модели със смесени ефекти 13

3 Модел за една наредена категорна променлива с повторни наблюдения 16

3.1 Описание на модела 16

3.2 Намиране на максимално правдоподобни оценки (МПО) чрез ЕМ алгоритъм 18

3.2.1 Лог-правдоподобие на пълния набор от данни 19
3.2.2 Явен вид на МПО 20
3.2.3 Условни очаквания 21
3.2.4 $(k+1)$-ва итерация на ECM алгоритъма 24
3.2.5 Приближение на стандартните грешки 25
3.3 Симулации 26
3.4 Приложение на модела 28
3.5 Заключение 30

4 Съвместен модел за няколко наредени категорни променливи 31
4.1 Описание на модела 31
4.2 Намиране на МПО чрез EM алгоритъм 32
4.2.1 Лог-правдоподобие на пълния набор от данни 33
4.2.2 Явен вид на МПО 34
4.2.3 Условни очаквания 35
4.2.4 $(k+1)$-ва итерация на ECM алгоритъма 37
4.2.5 Приближение на стандартните грешки 37
4.3 Симулации 38
4.4 Приложение на модела 40
4.5 Заключение 47

5 Съвместен модел за една наредена категорна и една нормална променливи, проследени неколкократно във времето 48
5.1 Описание на модела 48
5.2 МПО чрез EM алгоритъм 50
5.2.1 Лог-правдоподобие на пълния набор от данни 50
5.2.2 Явен вид на МПО 51
5.2.3 Условни очаквания 53
5.2.4 $(k+1)$-ва итерация на ECM алгоритъма 55
5.2.5 Приближение на стандартните грешки 56
5.3 Симулации 57
5.4 Приложение на модела 59
5.5 Заключение 62

6 Заключение **64**
6.1 Научен и практически принос на дисертацията 64
6.2 Бъдещи насоки за развитие 64

Публикувани и докладвани резултати, изложени в дисертационния труд **67**

Приложение А: R код за ECM алгоритъм за корелиран "probit" модел за една наредена категорна величина от дългосрочни проучвания **69**

Приложение Б: R код за ECM алгоритъм за корелиран "probit" модел за две наредени категорни величини **78**

Приложение В: R код за ECM алгоритъм за корелиран "probit" модел за една наредена категорна и една нормална величини от дългосрочни проучвания **86**

Библиография **99**

Декларация за оригиналност на резултатите **106**

Списък на фигурите

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

3.1 Нормално разпределена скрита променлива с 4 прага, която генерира наблюдавана категорна променлива с 5 нива. 17

Списък на таблиците

3.1 Оценки и стандартни грешки от двете симулационни изследвания за "probit" модела 3.3 за повторни наблюдения над една наредена категорна променлива 28
3.2 Таблица с оценки и стандартни грешки от модела 3.4, приложен към данните от здравно и пенсионно проучване 29

4.1 Оценки и стандартни грешки от двете симулационни изследвания за "probit" модела 4.3 за две наредени категорни величини 39
4.2 Съвместна таблица на кожни и урогенитални реакции от раково и генетично проучване 41
4.3 Таблица за разпределението на кожни реакции в зависимост от генотипа и за разпределението на урогенитални реакции в зависимост от генотипа 41
4.4 Таблица с оценки и стандартни грешки от съвместен модел (4.4) и отделни модели (4.5, 4.6) за кожни и урогенитални реакции от раково и генетично проучване 42
4.5 Таблица със съвместните вероятности за степента на кожните и урогениталните реакции според модела 4.4 в зависимост от генотипа на XRCC3 241 43
4.6 Таблица с вероятностите за степента на кожните и урогениталните реакции при отделно моделиране в зависимост от генотипа на XRCC3 241 44

5.1 Таблица с оценките и стандартните грешки от двете симулационни изследвания за “probit” модела 5.3 за повторни наблюдения на една наредена категорна и една нормална величини . . 59
5.2 Таблица с оценки и стандартни грешки от съвместния корелиран “probit” модел 5.4 за самооценка за здравето и индекса на телесна маса, приложен към данните от здравно и пенсионно проучване. 60
5.3 Таблица с оценки и стандартни грешки от корелирания “probit” модел 5.5 за самооценка за здравето, приложен към данните от здравно и пенсионно проучване. 61
5.4 Таблица с оценки и стандартни грешки от корелирания “probit” модел 5.6 за индекса на телесна маса, приложен към данните от здравно и пенсионно проучване. 62

Означения и съкращения

Означения

I_{n_i} единична матрица с размерност n_i

$\phi_{\boldsymbol{\mu},\,\boldsymbol{\Sigma}}(\boldsymbol{x})$ плътност на нормално разпределение със средно $\boldsymbol{\mu}$ и ковариационна матрица $\boldsymbol{\Sigma}$ в точката $\boldsymbol{x}$

$\circ$ поелементно умножение (умножение на Адамар)

$\otimes$ умножение на Кронекер

Съкращения

англ. английски

ЗПП здравно и пенсионно проучване

МПО максимално правдоподобна оценка

МНМК метод на най-малките квадрати

ОЛМ обобщен линеен модел

РГП раково и генетично проучване

сл. вел. случайна величина

стд. гр. стандартна грешка

стд. откл. стандартно отклонение

т. нар. така наречения

Глава 1

Въведение

"Всяко начало е трудно, но това не трябва
да ни спира да започваме нови неща!"

В следствие от развитието на технологиите е улеснено събирането, съхраняването и обработването на много по обем информация. В днешно време се провеждат много изследвания, които неколкократно във времето проследяват обектите от интерес. Целта на дългосрочните изследвания е да се установи влиянието на редица фактори, променящи се във времето, както и самото време, върху характеристики, които са от основен интерес за поръчителите на изследването. На статистиците се налага да развият нови статистически модели, които да опишат коректно и същевременно практически полезно сложни структури от данни, за които предположението за независимост е неприложимо. Първата задача, която възниква при новите модели, е оценяването на неизвестните параметри.

Данните с повторни наблюдения във времето (данни от дългосрочни изследвания, от англ. longitudinal data) обединяват характеристики от многомерни данни и времеви редове. Разликата от многомерни данни е, че съществува силна свързаност между последователните наблюдения във времето над един обект. Разликата от времевите редове е, че поредицата от наблюдения при данни с повторни наблюдения над един обект е много по-къса, отколкото е характерно за времевия ред.

Предимство на дългосрочния тип изследване е, че при него може да се

установят разликите в променливата от основен интерес във времето в рамките на всеки обект (т. нар. ефект на остаряване, от англ. aging effect) и разликите в променливата от основен интерес между обектите в момента на включването им в изследването (т. нар. групов ефект, от англ. cohort effect).

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 Diggle et al. [1996] дава широк, но по-теоретичен обзор върху изброените типове модели, докато тази на Weiss [2005] е подходяща и за по-незапознати със статистиката читатели. Книгата на Fitzmaurice et al. [2004] дава изключително изчерпателен поглед върху моделите, методите за оценяване и проблемите при работа с данни с повторни наблюдения.

При маргиналните модели се моделират очакването и ковариационната матрица на наблюденията в рамките на един обект. При оценяването на неизвестните параметри на модела се използва метод, наречен обобщени уравнения за оценяване (Liang and Zeger [1986]). При нормално разпределение на отклика моделирането на първите два момента напълно определя разпределението на променливата от основен интерес. Обобщените уравнения за оценяване дават оценки, които съвпадат с тези, получени при максимизиране на функцията на правдоподобие. При отсъствието на липсващи данни и коректен модел за средното получаваме състоятелни оценки за регресионните параметри дори при неточна структура за ковариационната матрица на наблюденията в рамките на един индивид.

Транзитивните (Марковски) модели са модели, при които предишните наблюдения над даден обект се използват като предиктори за моделирането на наблюдението в следващ момент.

При моделите със случайни ефекти (модели със смесени ефекти) (Laird and Ware [1982]) корелацията на наблюденията в рамките на един индивид

произлиза от общи за наблюденията на този индивид случайни величини. Познати са още като йерархични модели поради факта, че моделът може да се запише на две или повече нива. На първо ниво е моделът за математическото очакване на променливата от основен интерес условно по случайните ефекти. На второ ниво е моделът за разпределението на самите случайни ефекти. Най-често се прави предположение за нормално разпределение на случайните ефекти. Един подход за оценяването на параметрите в модела е максимизиране на маргиналното правдоподобие, което правдоподобие се получава след интегрирането на случайните ефекти. Моделите със случайни ефекти отразяват и факта, че индивидите имат различия в характеристики, които не се наблюдават директно и не могат да бъдат включени като предиктори в модела.

Моделирането на категорни величини е по-сложна задача от тази за моделирането на нормално разпределен отклик. 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). В книгата на Kutner et al. [2005] са разгледани случаите на "probit" и "logit" модели на една Бернулиева променлива, мерена по веднъж във времето, където латентната променлива има нормално и логистично разпределение съответно. Такива модели са наречени модели с праг, който подход датира от поне 1952 година (Edwards and Thurstone [1952]). 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 разпределение. Aitchison and Silvey [1957] предлагат "probit" модел за наредени категорни данни. Ochi and Prentice [1984] първи предлагат корелиран "probit" модел, но само за "exchangeable" бинарни данни. Разширения на тези

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

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 са разработени за корелирани бинарни и нормални данни от Gueorguieva and Agresti [2001], но не са разработени за корелирани ординални данни.

Стохастично оценяване е чрез EM алгоритми, които получават своето наименование от Dempster et al. [1977]. Dempster et al. [1977] развиват основната рамка на EM алгоритмите, които са предлагани от други автори преди тях, но в конкретни случаи. EM алгоритъмът е итеративна процедура за намиране на МПО на параметри в модели, които зависят от ненаблюдавани данни. Всяка итерация на алгоритъма се състои от две стъпки: E-стъпка (стъпка на очакването) и M-стъпка (стъпка на максимизирането). Едно разширение на EM алгоритмите е EM алгоритъмът с условно максимизиране в M-стъпката (от англ. Expectation/Conditional Maximization, Meng and Rubin [1993]) (ECM алгоритъм). Предимството на това разширение е заместването на сложната в някои случаи M-стъпка на алгоритъма с няколко изчислително по-прости CM-стъпки.

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. Тре-

тият 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. Такива предположения са правдоподобни, когато мерим характеристики, свързани със здравето на пациенти (физическо и психическо). 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 реализирането 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

Следващата глава Модели със случайни ефекти за данни с повторни наблюдения запознава читателя с основите на модели със случайни ефекти за данни с повторни наблюдения.

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, като използваният код на **R** е представен в Приложение A: R код за ECM алгоритъм за корелиран "probit" модел за една наредена категорна величина от дългосрочни проучвания.

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, чийто код е изложен в Приложе-

ние Б: R код за ECM алгоритъм за корелиран “probit” модел за две наредени категорни величини. С нея са извършени симулации, за да се изследва поведението на алгоритъма. С помощта на 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. Използваната за симулациите функция е представена в Приложение В: R код за ECM алгоритъм за корелиран “probit” модел за една наредена категорна и една нормална величини от дългосрочни проучвания.

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

Глава 2

Модели със случайни ефекти за данни с повторни наблюдения

"Once upon a time there were linear models with one normal error term." [1]

Моделите със смесени ефекти са подходящи за данни с повторни наблюдения или групирани данни. Те възникват от предположението, че данните могат да бъдат описани чрез йерархична структура. На първо място в тази йерархия стоят параметрите, които са общи за цялата популация. Това са фиксираните ефекти в модела. Допълнително всеки индивид има отличителни характеристики, които се описват чрез случайни ефекти. Това са случайни величини, за които се предполага някакво разпределение, най-често нормално.

Моделите със смесени ефекти добиват широко приложение след като статистиците се опитват да моделират по-прецизно дадена структура от данни и предположението за независимост на наблюденията е неадекватно.

Първо следва представянето на линеен модел с фиксирани ефекти и непосредствено след това линеен модел със случайни ефекти. След това са изложени обобщени линейни модели отново първо с фиксирани и след това със случайни ефекти.

[1] Това е първото изречение от увода на статията на Lee and Nelder [1996], стр. 619.

2.1 Линейни модели

2.1.1 Линейни модели с фиксирани ефекти

Линейните модели с фиксирани ефекти (Faraway [2004]; Kutner et al. [2005]) се записват по следния начин:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

където в извадката имаме n индивида и предполагаме, че грешките $\epsilon_i, i = 1, \ldots, n$ са независими, имат нормално разпределение със средно нула и в общия случай неизвестна дисперсия σ^2, която се оценява от данните. Откликът y_i за всеки индивид има нормално разпределение и очакването му е линейна функция на вектора от предикторите $\boldsymbol{x_i}$, които считаме за фиксирани (неслучайни). Векторът от неизвестните регресионни параметри е $\boldsymbol{\beta}$. Най-често оценките за $\boldsymbol{\beta}$ се намират по метода на най-малките квадрати (МНМК), които съвпадат с оценките по метода на максималното правдоподобие.

Коефициентите се интерпретират като средната промяна в отклика при единица промяна на съответната предикторна променлива.

2.1.2 Линейни модели със смесени ефекти

Променливата от основен интерес означаваме с y. Наблюдението над i-тия обект в момент j означаваме с y_{ij}. Целият вектор от наблюдения над обект i, който е с дължина n_i, е:

$$\boldsymbol{y_i} = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix}.$$

Векторът с времевите моменти на наблюденията над i-тия обект е:

$$\boldsymbol{t_i} = \begin{pmatrix} t_{i1} \\ t_{i2} \\ \vdots \\ t_{in_i} \end{pmatrix}.$$

При линейните модели със смесени ефекти (Diggle et al. [1996]; Weiss [2005]) моделираме многомерния отклик $\boldsymbol{y_i}$ като в стохастичното уравнение се добавя и част за случайните ефекти:

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

В извадката имаме n индивида с n_i наблюдения над всеки. Предполагаме, че грешките $\epsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i$ са независими помежду си и имат нормално разпределение $N(0, \sigma^2)$. Допускаме, че случайните ефекти $\boldsymbol{b_i}$ имат многомерно нормално разпределение $N(\boldsymbol{0}, \boldsymbol{\Sigma})$, независими за отделните индивиди и независими от грешките. Откликът y_{ij} за всеки индивид условно по случайните ефекти има нормално разпределение и очакването му е линейна функция на вектора от предикторите $\boldsymbol{x_{ij}}, j = 1, \ldots, n_i$, които считаме за фиксирани (неслучайни) и е линейна функция на вектора от предикторите за случайните ефекти $\boldsymbol{z_{ij}}$. От основен интерес са неизвестните регресионните параметри $\boldsymbol{\beta}$.

Линейните модели със смесени ефекти имат затворена форма на маргиналното правдоподобие. Маргиналното правдоподобие представлява маргиналната (безусловната по случайните ефекти) многомерната плътност на отклика.

От записа на модела следва, че многомерното наблюдение над един обект има многомерно нормално разпределение: $\boldsymbol{y_i} \sim N(\boldsymbol{X_i\beta}, \sigma^2\boldsymbol{I_{n_i}} + \boldsymbol{Z_i\Sigma Z'_i})$, където

$$\boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{z'_{i1}} \\ \boldsymbol{z'_{i2}} \\ \vdots \\ \boldsymbol{z'_{in_i}} \end{pmatrix}$$

и матрицата с предикторите за всички наблюдения над обекта i е $\boldsymbol{X_i}$, съставена от вектор-редове $\boldsymbol{x'_{ij}}, j = 1, \ldots, n_i$. Нека $\boldsymbol{V}$ е блок диагонална матрица с $n_i \times n_i$ размерни блокове $\sigma^2 \boldsymbol{I_{n_i}} + \boldsymbol{Z_i \Sigma Z'_i}, i = 1, \ldots, n$. Тогава

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V}),$$

където матрицата $\boldsymbol{X}$ е съставена от матриците $\boldsymbol{X_i}, i = 1, \ldots, n$:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \vdots \\ \boldsymbol{X_n} \end{pmatrix}.$$

Маргиналното лог-правдоподобието на наблюдаваните данни има следния явен вид:

$$\ln L(\beta, \sigma^2, \boldsymbol{\Sigma}) = -0.5[\ln |\boldsymbol{V}| + \boldsymbol{(y - X\beta)'V^{-1}(y - X\beta)}]$$

Съществуват два широко използвани метода за намиране на оценките на регресионните и на ковариационните параметри. Единият е методът на максималното правдоподобие. Оценките, получени по този метод, са състоятелни, асимптотично нормално разпределени и неизместени и асимптотично ефективни по Рао-Крамер (при прости извадки и някои твърде общи условия от аналитичен характер, Димитров и Янев [1998]).

Другият метод е модификация на метода на максималното правдоподобие. Той се нарича метод на ограниченото максимално правдоподобие (от англ. restricted maximum likelihood estimation, предложен за пръв път от Patterson and Thompson [1971]). Предимството на втория метод е, че дава неизместени оценки за ковариационните параметри, докато оценките за регресионните параметри са почти същите.

Коефициентите се интерпретират като средната промяна в отклика при единица промяна на съответната предикторна променлива както в рамките на индивида, така и на популационно ниво.

2.2 Обобщени линейни модели

2.2.1 Обобщени линейни модели с фиксирани ефекти

Първо ще въведем обобщени линейни модели (ОЛМ) (McCullagh and Nelder [1989]) за данни без повторни наблюдения. Те позволяват едно общо представяне на набор от статистически модели, сред които са линейната, логистичната и Поасоновата регресии.

В ОЛМ предполагаме, че всеки отклик y_i е реализация на случайна величина от разпределение от експоненциалното семейство, което се характеризира със следната плътност: $f(y_i; \theta_i, \phi) = \exp[\{y_i\theta_i - \psi(\theta_i)\}/\phi + c(y_i, \phi)]$. В частност експоненциалното семейство включва нормалното, биномното и Поасоновото разпределение.

В ОЛМ средното на зависимата променлива μ_i зависи от предсказващите променливи $\boldsymbol{x_i}$ чрез: $E(y_i) = \mu_i = g^{-1}(\boldsymbol{x_i}\boldsymbol{\beta})$, където $\boldsymbol{x_i}$ е векторът от предиктори, а g е гладка монотонна свързваща функция (от англ. link function). Неизвестните регресионни коефициенти са означени с $\boldsymbol{\beta}$. За тяхното оценяване най-често се използва методът на максималното правдоподобие.

В кой да е обобщен линеен модел регресионните параметри могат да бъдат оценени, решавайки следните уравнения:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' v_i^{-1}(y_i - \mu_i(\boldsymbol{\beta})) = 0,$$

където $v_i = \mathrm{Var}(y_i)$ и $S(\boldsymbol{\beta})$ е първата производна на логаритъма на функцията на правдоподобие (на англ. score function). Решението $\hat{\boldsymbol{\beta}}$, което е МПО, в общия случай няма явен вид. Една възможност за намиране на $\hat{\boldsymbol{\beta}}$ е чрез алгоритъма на претеглените най-малки квадрати (от англ. iteratively reweighted least squares).

Когато не правим предположение за конкретното разпределение на зависимата променлива, а само дефинираме първия и втория момент от него, и отново моделираме очакването като функция на предсказващи променливи, умножени по неизвестни коефициенти, можем да използваме записаното по-горе уравнение за оценки на неизвестните параметри. Този подход е известен

като максимизиране на псевдо-правдоподобието (от англ. quasi-likelihood).

Интерпретацията на коефициентите зависи от разпределението на зависимата променлива. В случая на Бернулиев отклик и свързваща "logit" функция даден регресионен коефициент представлява промяната на логаритъма на отношението на вероятността за успех и вероятността за неуспех (отношението на вероятността за успех и вероятността за неуспех ще наричаме залог за успех, от англ. odds ratio) с увеличаване на една единица на съответстващата за коефициента независима променлива. Увеличение на лог-отношението на шансовете означава увеличение на вероятността за успех и обратното. Намаление на лог-отношението на шансовете означава намаление на вероятността за успех. Знакът на регресионния коефициент определя връзката между съответната независима променлива и лог-отношението на вероятностите. Положителен знак говори за положителна връзка между предиктора и вероятността за успех, а знак минус - за отрицателна.

2.2.2 Обобщени линейни модели със смесени ефекти

Можем да дефинираме ОЛМ със смесени ефекти (Diggle et al. [1996]) по следния начин:

- при дадени случайни ефекти $\boldsymbol{b_i}$, наблюденията на зависимата променлива $y_{i1}, y_{i2}, \ldots, y_{in_i}$ са независими и следват обобщен линеен модел с плътност от експоненциалното семейство $f(y_{ij}|\boldsymbol{b_i}) = \exp[\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\}/\phi + c(y_{ij}, \phi)]$, за които са изпълнени условията, свързващи условните моменти и параметри от плътността: $\mu_{ij} = E(y_{ij}|\boldsymbol{b_i}) = \psi'(\theta_{ij})$ и $\nu_{ij} = \mathrm{Var}(y_{ij}|\boldsymbol{b_i}) = \psi''(\theta_{ij})\phi$, удовлетворявайки $h(\mu_{ij}) = \boldsymbol{x'_{ij}\beta} + \boldsymbol{z'_{ij}b_i}$ и $\nu_{ij} = \nu(\mu_{ij})\phi$, където h и ν са известни свързваща и дисперсионна функции съответно и $\boldsymbol{z_{ij}}$ е подмножество на $\boldsymbol{x_{ij}}$;

- случайните ефекти $\boldsymbol{b_i}, i = 1, \ldots, m$ са независими с еднакво многомерно разпределение F.

Маргиналното лог-правдоподобие има следния вид:

$$\ln L = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{b_i}) f(\boldsymbol{b_i}) \, \mathrm{d}\, \boldsymbol{b_i}.$$

Маргиналното правдоподобие няма явен вид освен в случая на нормално разпределен отклик. За намиране на оценките на неизвестните параметри се прилагат методи като числено интегриране, EM алгоритми или аналитични приближения.

Численото интегриране (Liu and Pierce [1994]) представлява намиране на приближение на интеграла като функция на неизвестните параметри и неговото максимизиране по отношение на неизвестните параметри (Pinheiro and Chao [2006]). Такъв тип подход изисква много изчисления при голяма размерност на областта на интегриране, т.е. при голяма размерност на случайните ефекти.

EM алгоритъмът (Dempster et al. [1977]) е итеративна процедура, състояща се от две стъпки: E-стъпка (стъпка на очакването) и M-стъпка (стъпка на максимизирането). За прилагане на алгоритъма е нужно да се реши спрямо неизвестните параметри системата уравнения от първите производни на лог-правдоподобието на пълния набор от данни, приравнени на нула. Пълният набор от данни представлява множеството от наблюдавани стойности за променливата от основен интерес и ненаблюдаваните случайни ефекти. Оценките за неизвестните параметри имат явен вид в случая на нормално разпределен отклик. В E-стъпката трябва да се изчислят условните очаквания на функциите от случайните ефекти, участващи във формулите на оценките, при условие наблюдаваните данни. В M-стъпката се пресмятат новите стойности на параметрите. Алгоритъмът започва със задаване на начални стойности за неизвестните параметри. Преминава към извършване на изчисленията в E-стъпката. След това се обновяват оценките в M-стъпката. Итеративно се повтаря цикълът E-стъпка и M-стъпка. Алгоритъмът спира при определени условия за сходимост като една възможност за спиращ критерий е разликата в стойностите на параметрите при съседни итерации да е по-малка от предварително определено малко число.

EM алгоритъмът може да се обобщи по следния начин:

1. Избираме начални стойности $\mathbf{\Gamma^{(0)}}$ за вектора от неизвестни параметри $\mathbf{\Gamma}$.

2. Увеличаваме брояча за номера на итерация k с 1.

3. Е-стъпка: изчисляваме $E[\ln L(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\Gamma})|\boldsymbol{y}, \boldsymbol{\Gamma}^{(\boldsymbol{k})}]$, където $\ln L(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\Gamma}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) + \sum_{i=1}^{n} \ln f(\boldsymbol{y_i}|\boldsymbol{b_i})$ е лог-правдоподобието на пълния набор от данни.

4. М-стъпка: Намираме следващата оценка на вектора от неизвестни параметри $\boldsymbol{\Gamma}^{(\boldsymbol{k+1})}$ като стойностите, които максимизират условното очакване от стъпка 3.

5. Повтаряме стъпки 2, 3 и 4, докато постигнем сходимост на редицата от оценки: $|\boldsymbol{\Gamma}^{(\boldsymbol{k+1})} - \boldsymbol{\Gamma}^{(\boldsymbol{k})}| < \epsilon$ за всеки елемент от вектора, където ϵ е предварително зададено малко число.

Аналитичното приближение е приближение на първите производни на маргиналното лог-правдоподобие, за да се избегне интегриране (Breslow and Clayton [1993]). Основната идея е да се използват условните моди, а не условните очаквания. Това е еквивалентно на приближаване на условното разпределение на случайните ефекти при условие данните с нормално разпределение със същите мода и кривина. Чрез използването на модите вместо средните, интегрирането е заменено с оптимизиране, което може да се вмъкне в М-стъпката. Оценките на параметрите, получени по този метод, са изместени за бинарни данни и ординални данни с малко на брой категории.

Предимство на ЕМ алгоритъма е, че изчислителната трудност на метода не се увеличава експоненциално с увеличаване на размерността на случайните ефекти. Недостатък е, че има възможност ЕМ алгоритъмът да сходи към локален максимум. Затова се налага да се задават различни начални стойности.

Коефициентите в обобщените линейни модели със смесени ефекти описват промяната на индивидуалния отклик при разлики на обяснителните променливи в рамките на индивида. В случая на Бернулиев отклик и свързваща "logit" функция даден регресионен коефициент представлява промяната на логаритъма на залога за успех с увеличаване на една единица на съответната предсказваща променлива за конкретен индивид.

Глава 3

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].

3.1 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:

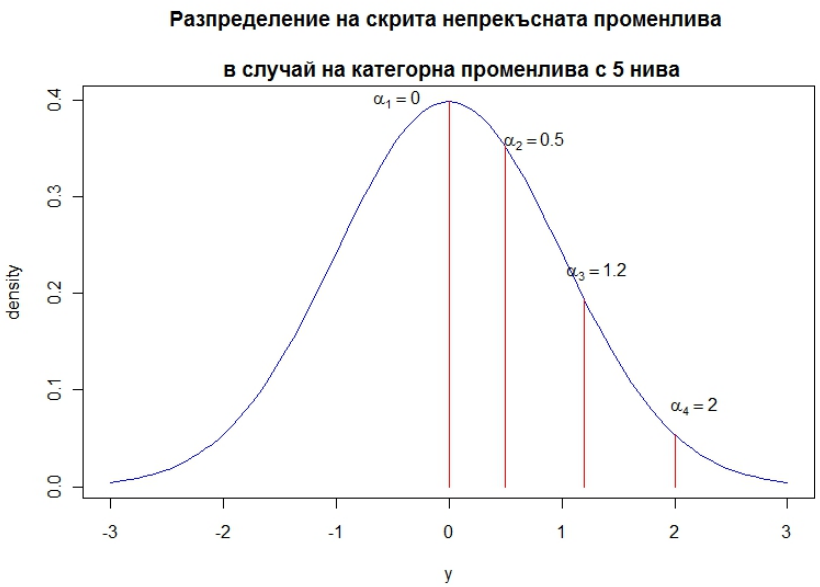
$$y_{ij} \quad = \quad \boldsymbol{x'_{ij}\beta} + \boldsymbol{z'_{ij}b_i} + \epsilon_{ij}. \tag{3.1}$$

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

$$y_{ij}^* = \begin{cases} 1, \ y_{ij} \leq \alpha_1; \\ l, \ \alpha_{l-1} < y_{ij} \leq \alpha_l, \ l = 2, \ldots, m-1; \\ m, \ y_{ij} > \alpha_{m-1}; \end{cases} \tag{3.2}$$

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

За илюстрация може да послужи Фигура 3.1. Ако скритата променлива е взела стойност между 0.5 и 1.2, то сме наблюдавали третото ниво на наредената категорна величина.



Фигура 3.1: Нормално разпределена скрита променлива с 4 прага, която генерира наблюдавана категорна променлива с 5 нива.

В модела 3.1 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 3.1 и неизвестните прагове в 3.2, 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.

В следващия раздел е представен EM алгоритъм (Dempster et al. [1977]) за намиране на оценки по метода на максимално правдоподобие.

3.2 Намиране на максимално правдоподобни оценки (МПО) чрез EM алгоритъм

Ние правим разширение на стохастичния ECM алгоритъм, предложен от Chan and Kuk [1997], за да оценим неизвестните параметри в модела 3.1 и неизвестните прагове в 3.2. 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 $\alpha_0 = 0$. Например, ако $y^*_{ij} = u,\ u = 1, \ldots, m$, тогава $y_{ij_{new}} = (y_{ij} - \alpha_{u-1})/\delta_u$. Понеже новата променлива е линейна трансформа-

ция на нормална сл. вел., то тя също има нормално разпределение. Тази променлива, условно по наблюдаваната категорна променлива, има орязано нормално разпределение и от дефиницията й 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 $(0, \infty)$. Да предположим, че $y^*_{ij} = u$ за някое $u \in \{2, \ldots, m-1\}$ и нека $y_{ij_{new}} \sim N(\mu^e_{ij}, \sigma^e_{ij})$. Тогава $y_{ij_{new}}$ при условие $y^*_{ij} = u$ има орязано нормално разпределение $N^T_{(0,1]}(\mu^e_{ij}, \sigma^e_{ij})$, където плътността на $y_{ij_{new}}|y^*_{ij} = u$ е $f(y_{ij_{new}}) = \frac{1/\sqrt{2\pi\sigma^e_{ij}}\exp(-(y_{ij_{new}}-\mu^e_{ij})^2/2\sigma^e_{ij})}{\int_0^1 1/\sqrt{2\pi\sigma^e_{ij}}\exp(-(x-\mu^e_{ij})^2/2\sigma^e_{ij})\,\mathrm{d}\,x} = \frac{\phi_{\mu^e_{ij},\sigma^e_{ij}}(y_{ij_{new}})}{P(y_{ij_{new}}\in(0,1])}$ за $y_{ij_{new}}$ в интервала $(0, 1]$ и 0 извън интервала $(0, 1]$, където $\phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x})$ е плътността на нормално разпределение със средно $\boldsymbol{\mu}$ и ковариационна матрица $\boldsymbol{\Sigma}$ в точката $\boldsymbol{x}$. Аналогично, променливата $\boldsymbol{y_{i_{new}}}|\boldsymbol{y^*_i}$ има n_i-мерно орязано нормално разпределение. Нека с A означим областта на орязване на $\boldsymbol{y_{i_{new}}}|\boldsymbol{y^*_i}$. Например, ако $\boldsymbol{y^*_i} = (u, m)$ за някое $u \in \{2, \ldots, m-1\}$, то областта на орязване е $(0, 1] \times (0, +\infty)$. Нека $\boldsymbol{y_{i_{new}}} \sim N(\boldsymbol{\mu^e_i}, \boldsymbol{\Sigma^e_i})$. Тогава $\boldsymbol{y_{i_{new}}}|\boldsymbol{y^*_i}$ има орязано нормално разпределение $N^T_A(\boldsymbol{\mu^e_i}, \boldsymbol{\Sigma^e_i})$, където плътността на $\boldsymbol{y_{i_{new}}}|\boldsymbol{y^*_i}$ е $f(\boldsymbol{y_{i_{new}}}) = \frac{\phi_{\boldsymbol{\mu^e_i},\boldsymbol{\Sigma^e_i}}(\boldsymbol{y_{i_{new}}})}{P(\boldsymbol{y_{i_{new}}}\in A)}$ за $\boldsymbol{y_{i_{new}}}$ в областта A и 0 извън областта A.

3.2.1 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 $\boldsymbol{y_{i_{new}}} = (y_{i1_{new}}, y_{i2_{new}}, \ldots, y_{in_{i_{new}}})', i = 1, \ldots, n$. За равенствата сме използвали факта, че случайните ефекти за не-

зависими помежду си и линейните трансформации на скритата величина в различните времеви моменти при условие случайните ефекти са независими.

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 $\boldsymbol{\delta} = (\delta_2, \ldots, \delta_{m-1})$.

3.2.2 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:

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

Следва, че регресионните параметри $\boldsymbol{\beta}$ са решение по МНМК при регресия на $\tilde{y}_{ij}$ по $\boldsymbol{x_{ij}}$, където $\tilde{y}_{ij} = \delta_{y_{ij}^*}y_{ij_{new}} - \boldsymbol{z_{ij}'}\boldsymbol{b_i} + \alpha_{y_{ij}^*-1}$.

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:

$$a = \sum_{i,j}\sum_{y_{ij}^*=k}(y_{ij_{new}}^2) + n_{k+1} + \ldots + n_m,$$

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

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. Това следва от факта, че константата a е положителна, а константата c е отрицателна.

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.

3.2.3 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:

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

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

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

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

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

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

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

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

Очакването на случайните ефекти при условие наблюдаваните данни е:

$$\begin{aligned}
E(\boldsymbol{b_i}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\
&= E[\boldsymbol{\Sigma_{B_i}}(\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}})|\boldsymbol{y_i^*}] \\
&= \boldsymbol{\Sigma_{B_i}}[\boldsymbol{E}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}].
\end{aligned}$$

За оценката на ковариационната матрица на случайните ефекти трябва да изчислим:

$$\begin{aligned}
E(\boldsymbol{b_i b_i'}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\
&= E[Var(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}}) + E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})E(\boldsymbol{b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma_{B_i}}[\boldsymbol{Z_i\Sigma} \circ (\boldsymbol{J_{n_i \times q}\delta_i^{-1}})] + \boldsymbol{\Sigma_{B_i}}E[\boldsymbol{M_i M_i'}|\boldsymbol{y_i^*}]\boldsymbol{\Sigma_{B_i}'} \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma_{B_i}}[\boldsymbol{Z_i\Sigma} \circ (\boldsymbol{J_{n_i \times q}\delta_i^{-1}})] + \\
&\quad \boldsymbol{\Sigma_{B_i}}[Var(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) + E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*})E(\boldsymbol{y_{i_{new}}'}|\boldsymbol{y_i^*}) \\
&\qquad -E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*})[(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]' \\
&\qquad -[(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]E(\boldsymbol{y_{i_{new}}'}|\boldsymbol{y_i^*}) \\
&\qquad +[(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}][(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]']\boldsymbol{\Sigma_{B_i}'}.
\end{aligned}$$

В израза за новите оценки на праговете участва следното условно очак-

ване:

$$
\begin{aligned}
E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_i^*}) &= E[E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\
&= E[y_{ij_{new}}\boldsymbol{\Sigma_{B_i}}(\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}})|\boldsymbol{y_i^*}] \\
&= \boldsymbol{\Sigma_{B_i}} E[y_{ij_{new}}\boldsymbol{y_{i_{new}}} - y_{ij_{new}}(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}|\boldsymbol{y_i^*}] \\
&= \boldsymbol{\Sigma_{B_i}}[Cov(y_{ij_{new}}, \boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) + E(y_{ij_{new}}|\boldsymbol{y_i^*})E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) \\
&\qquad -E(y_{ij_{new}}|\boldsymbol{y_i^*})(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}].
\end{aligned}
$$

Условните очаквания е възможно да бъдат изчислени с аналитични форми според работата на Manjunath and Wilhelm [2009], но изчисляването е трудно и времеемко особено, когато размерността на орязаното нормално разпределение е по-голяма от 2 (т.е., когато имаме повече от 2 наблюдения над обект) и голям брой обекти в извадката. Използваната апроксимация за реализацията на алгоритъма в Приложение A: R код за ECM алгоритъм за корелиран "probit" модел за една наредена категорна величина от дългосрочни проучвания е стохастична и дава стойности достатъчно близки до истинските. За да намерим условните очаквания, които показахме, че зависят само от първите два момента на $\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}$, сме използвали Монте Карло метод. Той се състои в генериране на числа от многомерно орязано нормално разпределение при условие наблюдаваните данни. За реализацията сме използвали метода за генериране на числа на Гибс (от англ. Gibbs sampling, Casella and George [1992]). Методът на Гибс е алгоритъм за получаване на редица от наблюдения, които са приближение на числа от многомерно вероятностно разпределение. Алгоритъмът се прилага в случаите, когато генерирането на числа от това разпределение е сложна задача.

Да предположим, че искаме да получим m на брой извадки от многомерното разпределение на $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ с функция на разпределение $F(x_1, x_2, \ldots, x_n)$. Нека да означим i-тата извадка с $\boldsymbol{x^{(i)}} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})'$. Стъпките на алгоритъма за генериране на числа на Гибс са следните:

1. Започваме с начална стойност $\boldsymbol{x^{(0)}}$.

2. За всяка извадка $i = 1, 2, \ldots, m$ генерираме стойността за $x_j^{(i)}$ от условното разпределение $F(x_j|x_1^{(i)}, x_2^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, x_n^{(i-1)})$.

Получените по този начин извадки са силно зависими. При голямо i алгоритъмът схожда към генериране на наблюдения от желаното разпределение и може да се приеме, че $\boldsymbol{x^{(i)}}$ е случаен вектор от търсеното вероятностно разпределение. Алгоритъмът е много по-бърз от този за генериране на числа чрез отхвърляне (от англ. rejection sampling), където наблюденията са независими. При метода на Гибс се практикува премахване на начално генерираните наблюдения, за да се осигури, че извадката е от търсеното разпределение. За преодоляване на проблема със зависимостта се практикува изтъняване на извадката - взима се всяко $p-$то генерирано наблюдение (например стотно, двестно).

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 $Var(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*})$. С математически означения го записваме по следния начин:

$$\begin{aligned}
\hat{E}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) &= \frac{1}{n_{mc}}\sum_{k=1}^{n_{mc}} \boldsymbol{y_{i_{new}}^{(k)}}, \\
\hat{Var}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) &= \frac{1}{n_{mc}-1}\sum_{k=1}^{n_{mc}} (\boldsymbol{y_{i_{new}}^{(k)}} - \boldsymbol{\hat{E}}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}))(\boldsymbol{y_{i_{new}}^{(k)}} - \boldsymbol{\hat{E}}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}))',
\end{aligned}$$

където $\boldsymbol{y_{i_{new}}^{(k)}}$ е $k-$тата реализация на $\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}$ в генерираната извадка от n_{mc} случайни числа. Според нашия опит дори за малко n_{mc} (например 150 или 200) се получават достатъчно близки до истинските резултати.

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

Ние използваме разширение на EM алгоритъма, наречено условно максимизиране в M-стъпката (Expectation/Conditional Maximisation algorithm (Meng and Rubin [1993])). Разликата на ECM алгоритъма от EM алгоритъма е, че в M-стъпката всеки параметър се максимизира индивидуално при условие, че останалите параметри стоят фиксирани. M-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 $E(\tilde{y}_{ij}|\boldsymbol{y_i^*;\Gamma^k})$ по $\boldsymbol{x_{ij}}$.

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

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

- $(k+1)$-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 $\boldsymbol{\beta^{k+1}}$, $\delta_i^{k+1}, i=2,\dots,m-1$.

За да пресметнем оценките, използваме приближени стойности за очакванията и дисперсиите, описани в предишната част.

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.

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. "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.

Стъпките за процедурата за намиране на приближение на стандартните грешки са следните:

1. Оценяваме параметрите $\boldsymbol{\Gamma}$ на модела 3.1 и връзката 3.2 според данните, състоящи се от n обекта, използвайки предложения ECM алгоритъм и получаваме оценки за неизвестните параметри, означени с $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\delta}})$. За да генерираме "bootstrap" извадка, първо генерираме n случайни ефекта $\boldsymbol{b_k^b}$ от $N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, $k = 1, \ldots, n$. След това симулираме нормални вектори $\boldsymbol{y_k^b}$ с размерност n_k според модел 3.1 за всеки случаен ефект $\boldsymbol{b_k^b}$. Използваме оценените чрез $\hat{\boldsymbol{\delta}}$ прагове, за да определим в кой интервал попадат нормалните данни $\boldsymbol{y_k^b}$, $k = 1, \ldots, n$ и използвайки връзката между скритата и наблюдаваната величина 3.2, определяме нивото на "bootstrap" категорната променлива $\boldsymbol{y_k^{b*}}$. Така получаваме "bootstrap" извадка, състояща се от категорните променливи $\boldsymbol{y_k^{b*}}$, $k = 1, \ldots, n$.

2. Прилагаме ECM алгоритъма към получените "bootstrap" данни $\boldsymbol{y_k^{b*}}$, $k = 1, \ldots, n$, за да получим оценки $\boldsymbol{\Gamma^b}$ за генерираното "bootstrap" множество от данни.

3. Използваме Монте Карло метод, за да намерим приближение на "bootstrap" ковариационната матрица. Повтаряме стъпка 1 и стъпка 2 B пъти и пресмятаме ковариационната матрица на B оценени параметри $\boldsymbol{\Gamma^b}, b = 1, \ldots, B$:
$$Cov(\hat{\boldsymbol{\Gamma}}) \approx \sum_{b=1}^{B} \frac{(\boldsymbol{\Gamma^b} - \bar{\boldsymbol{\Gamma}})(\boldsymbol{\Gamma^b} - \bar{\boldsymbol{\Gamma}})'}{B-1},$$
където $\bar{\boldsymbol{\Gamma}} = \sum_{b=1}^{B} \boldsymbol{\Gamma^b}/B$.

3.3 Симулации

За реализирането на алгоритъма е използвана безплатната софтуерна среда за статистически изчисления и графики **R** (R Core Team [2013]). Кодът на **R** за функцията, изпълняваща описания по-горе ECM алгоритъм е представен в Приложение A: R код за ECM алгоритъм за корелиран "probit" модел за една наредена категорна величина от дългосрочни проучвания. За оценяването на първите два момента на орязано нормално разпределение сме използвали

извадки, получени по метода за генериране на числа на Гибс (Casella and George [1992]) чрез функцията **rtmvnorm** в средата **R** от пакета **tmvtnorm** (Wilhelm and Manjunath [2012]). Точният алгоритъм за генериране на числата е описан в детайли в работата на Wilhelm [2012].

Симулирани са стойности от следния модел със случаен свободен член:

$$y_{ij} \quad = \quad \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij}, \; j = 1, \ldots, 5, \tag{3.3}$$

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 3.1.

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.

Както се очаква от теоретичните резултати за МПО, оценките са асимптотично неизместени и стандартните грешки намаляват с увеличаване на големината на извадката. Всички оценки са статистически значимо отличими от нула, освен дисперсията на случайния свободен член за по-малката извадка. Този резултат не е изненадващ, тъй като истинската стойност на дисперсията е много малка и оценките при малка големина на извадката не са толкова ефективни, колкото при голяма извадка.

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.

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

3.4 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:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{ij}, \\ \text{srh}_{ij} &= \begin{cases} 1, \ y_{ij} \leq \alpha_1 = 0, \\ l, \ \alpha_{l-1} < y_{ij} \leq \alpha_l, \ l = 2,3,4, \\ 5, \ y_{ij} > \alpha_4, \end{cases} \end{aligned} \tag{3.4}$$

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 3.2.

Таблица 3.2: Таблица с оценки и стандартни грешки от модела 3.4, 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

Таблица 3.2 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 $p - value < 0.0001$. Следователно 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). Т.к. тестваме дали параметър е на границата на допустимото множество, то разпределението на тест-статистиката не е хи-квадрат с една степен на свобода, а е смес от хи-квадрат разпределения (Weiss [2005]). Веро-

ятностната стойност на теста, че дисперсията на случайния свободен член е нула срещу алтернативата, че е по-голяма от нула, е по-малка от 0.0001 и 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.

3.5 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.

Възможно е разширение на алгоритъма с цел подобряване на бързината на схождане чрез увеличаване броя на параметрите, които се оценяват (Liu et al. [1998]). Може например към множеството на оценяваните параметри да се добави дисперсията на грешките (която за пълна определеност на модела ограничихме да е 1). При такъв подход в последната итерация на алгоритъма стойностите на останалите параметри трябва да се разделят на корен квадратен от оценката на дисперсията на грешките и така отново ще се получи еднозначна определеност на параметрите, задаващи модела.

Използваният метод за изчисление на стандартните грешки е изключително времеемък и тежък от изчислителна гледна точка, но при него няма риск да се получат стойности извън допустимото множество. Възможно подобрение по отношение на двата отбелязани недостатъка е разработването на функция за оценяване на стандартните грешки чрез приближение на Луис (Louis [1982]).

Следващата глава представя разширение на предложения ECM алгоритъм за модел на няколко категорни променливи с наредба на нивата.

Глава 4

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 Модел за една наредена категорна променлива с повторни наблюдения. 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.

4.1 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:

$$y_{ij} = \boldsymbol{x'_{ij}\beta_j} + \boldsymbol{z'_{ij}b_{ij}} + \epsilon_{ij},\ j = 1, \ldots, p, \quad \text{където сме наблюдавали} \tag{4.1}$$

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

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

Предполагаме, че векторът от случайните ефекти $\boldsymbol{b_i} = \boldsymbol{(b_{i1}, b_{i2}, \ldots, b_{ip})'}$ 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 4.1 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 $\boldsymbol{b} = \boldsymbol{(b_1, b_2, \ldots, b_n)'}$ 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.

Оценяването на всички параметри на модела и всички прагове води до нееднозначност. 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. Това не е единствената възможност за пълна определеност на модела.

4.2 Намиране на МПО чрез ЕМ алгоритъм

Разширяваме разработения в предишната глава ЕСМ алгоритъм, за да оценим неизвестните параметри в модела 4.1 и неизвестните прагове в 4.2. Началната стъпка е да въведем праговете във функцията на пълното правдо-

подобие (използваме подхода на Kawakatsu and Largey [2009] и Ruud [1991]). Дефинираме разликите между съседните прагове с $\delta_{j,i} = \alpha_{j,i} - \alpha_{j,i-1},\ i = 2,\dots,m_j - 1, j = 1,\dots,p$. Оттук следва връзката, че $\alpha_{j,i} = \sum_{k=2}^{i} \delta_{j,k},\ j = 1,\dots,p,\ i = 2,\dots,m_j - 1$. Полагаме $\delta_{j,1} = \delta_{j,m_j} = 1$ и $\alpha_{j,0} = 0, j = 1,\dots,p$. Следващата стъпка е да разгледаме нови променливи, които са линейна трансформация на скритите величини: $y_{ij_{new}} = (y_{ij} - \alpha_{j,y^*_{ij}-1})/\delta_{j,y^*_{ij}}, j = 1,\dots,p$ и нека $\boldsymbol{y_{i_{new}}} = (y_{i1_{new}}, y_{i2_{new}},\dots,y_{ip_{new}})'$. Например, ако $y^*_{ij} = u,\ u = 1,\dots,m_j$, тогава $y_{ij_{new}} = (y_{ij} - \alpha_{j,u-1})/\delta_{j,u}, j = 1,\dots,p$. Понеже новите променливи са линейни трансформации на нормални сл. вел., то те също имат нормално разпределение. Променливата $\boldsymbol{y_{i_{new}}}$, условно по наблюдаваните категорни променливи $\boldsymbol{y^*_{i_{new}}}$, има $p-$мерно орязано нормално разпределение.

4.2.1 Лог-правдоподобие на пълния набор от данни

Пълното лог-правдоподобие $\ln L$ е плътността на многомерното разпределение на случайните ефекти и линейните трансформации на скритите нормални величини в стойностите, които бихме наблюдавали:

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

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

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

$$\begin{aligned}
\ln L \;=\; & -0.5\sum_{i=1}^{n} \ln|\boldsymbol{\Sigma}| - 0.5\sum_{i=1}^{n} \boldsymbol{b'_i}\boldsymbol{\Sigma^{-1}}\boldsymbol{b_i} + \\
& + \sum_{i=1}^{n} \ln \delta_{1,y^*_{i1}} - 0.5\sum_{i=1}^{n} [\delta_{1,y^*_{i1}} y_{i1_{new}} - (\boldsymbol{x'_{i1}}\boldsymbol{\beta_1} + \boldsymbol{z'_{i1}}\boldsymbol{b_{i1}} - \alpha_{1,y^*_{i1}-1})]^2 \\
& + \sum_{i=1}^{n} \ln \delta_{2,y^*_{i2}} - 0.5\sum_{i=1}^{n} [\delta_{2,y^*_{i2}} y_{i2_{new}} - (\boldsymbol{x'_{i2}}\boldsymbol{\beta_2} + \boldsymbol{z'_{i2}}\boldsymbol{b_{i2}} - \alpha_{2,y^*_{i2}-1})]^2 \\
& + \dots \\
& + \sum_{i=1}^{n} \ln \delta_{p,y^*_{ip}} - 0.5\sum_{i=1}^{n} [\delta_{p,y^*_{ip}} y_{ip_{new}} - (\boldsymbol{x'_{ip}}\boldsymbol{\beta_p} + \boldsymbol{z'_{ip}}\boldsymbol{b_{ip}} - \alpha_{p,y^*_{ip}-1})]^2.
\end{aligned}$$

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 $\boldsymbol{\delta_j} = (\delta_{j,2}, \ldots, \delta_{j,m_j-1})$, $j = 1, \ldots, p$.

4.2.2 Явен вид на МПО

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

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

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

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

Следва, че регресионните параметри $\boldsymbol{\beta_j}, j = 1, \ldots, p$ са решение по МНМК при регресия на $\tilde{y}_{ij}$ по $\boldsymbol{x_{ij}}$, където $\tilde{y}_{ij} = \delta_{j,y_{ij}^*} y_{ij_{new}} - \boldsymbol{z_{ij}'}\boldsymbol{b_{ij}} + \alpha_{j,y_{ij}^*-1}, j = 1, \ldots, p$.

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

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

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 неизвестни-

те параметри. Изчисляването на новите оценки зависи от очакванията на сл.вел., участващи по-горе, при условие наблюдаваните данни. Следва параграф, който показва, че тези условни математически очаквания зависят само от първите два момента на $p-$мерно орязано нормално разпределение.

4.2.3 Условни очаквания

Нека използваме следните означения за опростяване на записа:

$$\boldsymbol{X_i} = \begin{pmatrix} \boldsymbol{x'_{i1}} & \boldsymbol{0'} & \dots & \boldsymbol{0'} \\ \boldsymbol{0'} & \boldsymbol{x'_{i2}} & \dots & \boldsymbol{0'} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0'} & \boldsymbol{0'} & \dots & \boldsymbol{x'_{ip}} \end{pmatrix}, \ \boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{z'_{i1}} & \boldsymbol{0'} & \dots & \boldsymbol{0'} \\ \boldsymbol{0'} & \boldsymbol{z'_{i2}} & \dots & \boldsymbol{0'} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0'} & \boldsymbol{0'} & \dots & \boldsymbol{z'_{ip}} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_p} \end{pmatrix},$$

$$\boldsymbol{\alpha_i} = \begin{pmatrix} \alpha_{1,y^*_{i1}-1} \\ \alpha_{2,y^*_{i2}-1} \\ \vdots \\ \alpha_{p,y^*_{ip}-1} \end{pmatrix}, \ \boldsymbol{\delta_i^{-1}} = \begin{pmatrix} 1/\delta_{1,y_{i1}} \\ 1/\delta_{2,y_{i2}} \\ \vdots \\ 1/\delta_{p,y_{ip}} \end{pmatrix}.$$

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

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

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

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

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

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

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

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

В изразите на оценките на Е-стъпката на алгоритъма участват следните

условни очаквания: $E(\boldsymbol{b_i}|\boldsymbol{y_i^*}), E(\boldsymbol{b_i}\boldsymbol{b_i'}|\boldsymbol{y_i^*}), E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_i^*})$. Ще покажем, че те зависят само от първите два момента на $\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}$, което е $p-$мерно орязано нормално разпределение.

Очакването на случайните ефекти при условие наблюдаваните данни е:

$$\begin{aligned} E(\boldsymbol{b_i}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E[\boldsymbol{\Sigma_{B_i}}(\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}})|\boldsymbol{y_i^*}] \\ &= \boldsymbol{\Sigma_{B_i}}[\boldsymbol{E}(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]. \end{aligned}$$

За оценката на ковариационната матрица на случайните ефекти трябва да изчислим:

$$\begin{aligned} E(\boldsymbol{b_i}\boldsymbol{b_i'}|\boldsymbol{y_i^*}) &= E[E(\boldsymbol{b_i}\boldsymbol{b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E[Var(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}}) + E(\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})E(\boldsymbol{b_i'}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma_{B_i}}(\boldsymbol{Z_i\Sigma} \circ (\boldsymbol{J_{p\times q}}\boldsymbol{\delta_i^{-1}})) + \boldsymbol{\Sigma_{B_i}}E[\boldsymbol{M_iM_i'}|\boldsymbol{y_i^*}]\boldsymbol{\Sigma'_{B_i}} \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma_{B_i}}(\boldsymbol{Z_i\Sigma} \circ (\boldsymbol{J_{p\times q}}\boldsymbol{\delta_i^{-1}})) + \\ &\quad \boldsymbol{\Sigma_{B_i}}[Var(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) + E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*})E(\boldsymbol{y'_{i_{new}}}|\boldsymbol{y_i^*}) \\ &\qquad -E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*})[(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]' \\ &\qquad -[(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]E(\boldsymbol{y'_{i_{new}}}|\boldsymbol{y_i^*}) \\ &\qquad +[(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}][(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]']\boldsymbol{\Sigma'_{B_i}}. \end{aligned}$$

В израза за новите оценки на праговете участва следното условно очакване:

$$\begin{aligned} E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_i^*}) &= E[E(y_{ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_{i_{new}}})|\boldsymbol{y_i^*}] \\ &= E[y_{ij_{new}}\boldsymbol{\Sigma_{B_i}}(\boldsymbol{y_{i_{new}}} - (\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}})|\boldsymbol{y_i^*}] \\ &= \boldsymbol{\Sigma_{B_i}}E[y_{ij_{new}}\boldsymbol{y_{i_{new}}} - y_{ij_{new}}(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}|\boldsymbol{y_i^*}] \\ &= \boldsymbol{\Sigma_{B_i}}[Cov(y_{ij_{new}}\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) + E(y_{ij_{new}}|\boldsymbol{y_i^*})E(\boldsymbol{y_{i_{new}}}|\boldsymbol{y_i^*}) \\ &\qquad -E(y_{ij_{new}}|\boldsymbol{y_i^*})(\boldsymbol{X_i\beta} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}}]. \end{aligned}$$

4.2.4 $(k+1)$-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, както в предишната глава, е разликата в стойностите на параметрите в две последователни итерации да е по-малка от предварително зададено 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. Стъпките са следните:

1. Използвайки предложения по-горе ECM алгоритъм, оценяваме неизвестните параметри в модела 4.1 и неизвестните прагове в 4.2 от данните, които се състоят от n обекта. Получаваме оценки за неизвестните параметри, означени с $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_p, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\delta}}_p)$. За да генерираме "bootstrap" извадка, първо генерираме n случайни ефекта $\boldsymbol{b_k^b}$ от $N(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}})$, $k = 1, \ldots, n$. След това симулираме $p-$мерни нормални вектори $\boldsymbol{y_k^b}$ според модел 4.1 при условие случаен ефект $\boldsymbol{b_k^b}$. Използваме оценените прагове $\hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\delta}}_p$, за да определим в кой интервал попадат нормалните данни $\boldsymbol{y_k^b}$, $k = 1, \ldots, n$ и използвайки връзката

4.2 определяме нивата на $p-$мерната "bootstrap" категорна променлива $\boldsymbol{y_k^{b*}}$. Така получаваме "bootstrap" извадка, състояща се от категорните променливи $\boldsymbol{y_k^{b*}}$, $k = 1, \ldots, n$.

2. Прилагаме ECM алгоритъма към получените "bootstrap" данни $\boldsymbol{y_k^{b*}}$, $k = 1, \ldots, n$, за да получим оценки $\boldsymbol{\Gamma^b}$ за генерираното множество от данни.

3. Използваме Монте Карло метод, за да намерим приближение на "bootstrap" ковариационната матрица. Повтаряме стъпка 1 и стъпка 2 B пъти и пресмятаме ковариационната матрица на B оценени параметри $\boldsymbol{\Gamma^b}, b = 1, \ldots, B$:

$$Cov(\hat{\boldsymbol{\Gamma}}) \approx \sum_{b=1}^{B} \frac{(\boldsymbol{\Gamma^b} - \bar{\boldsymbol{\Gamma}})(\boldsymbol{\Gamma^b} - \bar{\boldsymbol{\Gamma}})'}{B-1},$$

където $\bar{\boldsymbol{\Gamma}} = \sum_{b=1}^{B} \boldsymbol{\Gamma^b}/B$.

4.3 Симулации

За реализирането на алгоритъма сме използвали безплатната софтуерна среда за статистически изчисления и графики **R**. Кодът на функцията, когато размерността за отклика е две, е представен в Приложение Б: R код за ECM алгоритъм за корелиран "probit" модел за две наредени категорни величини. Условните очаквания, необходими за E-стъпката на алгоритъма, е възможно да бъдат изчислени с аналитични форми според работата на Manjunath and Wilhelm [2009]. За да пресметнем оценките, използваме истинските стойности за очакванията и дисперсиите, изчислени с помощта на функцията **mtmvnorm** в пакета **tmvtnorm** (Wilhelm and Manjunath [2012]). При реализацията на алгоритъма сме използвали 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:

$$y_{i1} \quad = \quad \beta_{10} + \beta_{11} x_{i1} + b_{i1} + \epsilon_{i1},$$

$$y_{i2} \quad = \quad \beta_{20} + \beta_{21} x_{i2} + b_{i2} + \epsilon_{i2}, \tag{4.3}$$

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, което е в границите на препоръчителните от 50 до 100 "bootstrap" повторения (Efron and Tibshirani [1994]). Резултатите са представени в Таблица 4.1.

Таблица 4.1: Оценки и стандартни грешки от двете симулационни изследвания за "probit" модела 4.3 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

Да отбележим, че поради репараметризацията, ние оценяваме разликите в праговете, а не самите тях. Но в случая на само три нива на категорните променливи, разликата между втория и първия праг (който е нула) съвпада със стойността на втория праг. 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. Всички оценки са статистически значимо отличими от нула с изключения на свободния член за първата променлива при по-малката извадка.

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.

4.4 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 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 урогенитални-те 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. Той е отговорен за откриването и поправя-

нето на увредени ДНК молекули. Медиците са заинтересувани доколко той е свързан с проявата на странични реакции след лъчелечение. Променливата XRCC3 241 взима стойност 0 за генотип {C,C} (45 наблюдения) и 1 за генотип {C,T} или {T,T} (76 наблюдения). Индивидите с генотип {C,T} или {T,T} са обединени в една категория, т.к. има твърде малко на брой пациенти с генотип {T,T} като това решение е съгласувано с медицинските лица. Обобщено представяне на данните има в Таблица 4.2 и Таблица 4.3.

Таблица 4.2: Съвместна таблица на кожни и урогенитални реакции от раково и генетично проучване

	Урогенитални реакции		
Кожни реакции	отсъстващи	слаби	умерени или силни
отсъстващи	24	17	7
слаби	12	16	14
умерени или силни	7	12	12

Таблица 4.3: Таблица за разпределението на кожни реакции в зависимост от генотипа и за разпределението на урогенитални реакции в зависимост от генотипа

	Кожни реакции		
XRCC3 241	отсъстващи	слаби	умерени или силни
{C,C}	11	19	15
{C,T} или {T,T}	37	23	16
	Урогенитални реакции		
XRCC3 241	отсъстващи	слаби	умерени или силни
{C,C}	16	17	12
{C,T} или {T,T}	27	28	21

Приложихме следния съвместен модел за кожни и урогенитални реакции от РГП:

$$
\begin{aligned}
y_{ij} &= \beta_{j0} + \beta_{j1}\text{XRCC3 241}_i + b_{ij} + \epsilon_{ij}, \qquad (4.4)\\
\text{Реакции}_{ij} &= \begin{cases} \text{Отсъстващи}, y_{ij} \leq \alpha_{j,1}, (\alpha_{j,1} = 0), \\ \text{Слаби},\ 0 < y_{ij} \leq \alpha_{j,2}, \\ \text{Умерени или силни},\ y_{ij} > \alpha_{j,2}, \end{cases}
\end{aligned}
$$

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 4.4: Таблица с оценки и стандартни грешки от съвместен модел (4.4) и отделни модели (4.5, 4.6) 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 4.4									
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 4.5									
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 4.6									
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 реакции. Моделът за кожни реакции е следният:

$$\begin{aligned} y_i &= \beta_{10} + \beta_{11}\text{XRCC3 241}_i + \epsilon_i, \qquad (4.5) \\ \text{Кожни реакции}_i &= \begin{cases} \text{Отсъстващи}, y_i \leq \alpha_{1,1}, (\alpha_{1,1} = 0), \\ \text{Слаби}, \ 0 < y_i \leq \alpha_{1,2}, \\ \text{Умерени или силни}, \ y_i > \alpha_{1,2}, \end{cases} \end{aligned}$$

където $\epsilon_i \sim N(0, 1)$.

Моделът за урогенитални реакции е аналогичен:

$$y_i = \beta_{20} + \beta_{21}\text{XRCC3 241}_i + \epsilon_i, \qquad (4.6)$$

$$\text{Урогенитални реакции}_i \quad = \quad \begin{cases} \text{Отсъстващи}, y_i \leq \alpha_{2,1}, (\alpha_{2,1} = 0), \\ \text{Слаби}, \ 0 < y_i \leq \alpha_{2,2}, \\ \text{Умерени или силни}, \ y_i > \alpha_{2,2}, \end{cases}$$

където $\epsilon_i \sim N(0, 1)$.

Резултатите от трите модела са представени в Таблица с оценки и стандартни грешки от съвместен модел (4.4) и отделни модели (4.5, 4.6) за кожни и урогенитални реакции от раково и генетично проучване. Таблица 4.4 показва, че всички параметри в съвместния модел 4.4 са статистически значимо отличими от нула освен индикаторната променлива за генотипа на полиморфизъм XRCC3 кодон 241 (C>T) в подмодела за урогениталните реакции β_{21}.

Таблица 4.5: Таблица със съвместните вероятности за степента на кожните и урогениталните реакции според модела 4.4 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 $z = -0.772/0.264 = -2.928$, 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

Таблица 4.6: Таблица с вероятностите за степента на кожните и урогенитал-ните 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

заключим, че генотип {C,C} на полиморфизъм XRCC3 241 повишава риска от по-тежки кожни реакции. Тези резултати са близки до датско следмастектомия изследване, при което е открита значима връзка между дивия алел C на кодон 241 и повишения риск от подкожна фиброза в група от 41 пациента. В същото изследване е открита и аналогична връзка за поява на телеангиектазия (Andreassen et al. [2003]). При други изследвания, целящи да потвърдят тези резултати, не са открити такива връзки между тези едно-нуклеотидни полиморфизми и риска от късни странични реакции (Andreassen [2005]; De Ruyck et al. [2005]). По-нататъшни изследвания, включващи допълнителни променливи, могат да разкрият допълнително зависимости между степента на странични реакции и други характеристики на индивидите.

Вероятностите за силата на кожните и урогениталните реакции според модела 4.4 в зависимост от генотипа на пациентите са представени в Таблица 4.5. 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 потискащ реакциите ефект.

Таблица 4.6 представя вероятностите за силата на кожните и урогенитал-

ните 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 4.5. 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 4.4 нямаме такива усложнения. Например, ако искаме да тестваме хипотезата, че генотипът на полиморфизъм XRCC3 241 има ефект върху проявата на кожни или урогенитални реакции:

$$\begin{aligned} H_0 : &\quad \beta_{11} = \beta_{21} = 0 \\ H_a : &\quad \text{поне едно от } \beta_{11} \text{ или } \beta_{21} \text{ е различно от } 0, \end{aligned}$$

можем да използваме теста на Уолд. Според оценките, получени от модел 4.4, стойността на тест-статистиката е $\chi^2_{df=2} = 8.85$, p-value=0.012. При ниво на значимост $\alpha = 0.05$ заключенията от теста са, че генотипът на полиморфизъм XRCC3 241 има ефект върху проявата на кожни или урогенитални реакции.

При тестване на параметрите при отделните модели: за кожни реакции тестваме:

$$\begin{aligned} H_0 : &\quad \beta_{11} = 0 \\ H_a : &\quad \beta_{11} \neq 0 \end{aligned}$$

и получаваме стойност на тест-статистиката $\chi^2_{df=1} = 6.04$, p-value=0.014; за

урогенитални реакции тестваме

$$H_0: \quad \beta_{21} = 0$$
$$H_a: \quad \beta_{21} \neq 0$$

и получаваме стойност на тест-статистиката $\chi^2_{df=1} = 0.004$, p-value=0.947. Ако използваме корекция при отделните тестове от отделните модели ще сравняваме вероятностните стойности с коригирано ниво на значимост, в случая на Бонферони $\alpha/2 = 0.025$. Тогава ще отхвърлим нулевата хипотеза, че генотипът на полиморфизъм XRCC3 241 има ефект върху проявата на кожни странични реакции при едновременно тестване на двете хипотези, че генотипът на полиморфизъм XRCC3 241 има ефект върху проявата на кожни и урогенитални странични реакции.

В конкретния случай и двата подхода на тестване водят до едни и същи заключения. Но би могло да се случи всички вероятностни стойности да бъдат над коригираното ниво на значимост при отделните тестове, докато при теста за тези параметри при съвместния модел вероятностната стойност да е под общото ниво на значимост. Това ще доведе до различни статистически заключения. Тогава възниква въпросът каква корекция е удачно да използваме и доколко можем да разчитаме на направените на база корекцията заключения.

Също така дисперсиите на случайните свободни членове според модел 4.4 са статистически значимо по-големи от 0 (z-score$_{\sigma_{11}} = 5.255$, z-score$_{\sigma_{22}} = 5.327$). Ковариацията между случайните ефекти е статистически значимо по-голяма от нула (z-score$_{\sigma_{12}} = 2.610$). От това можем да заключим, че зависимостта 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.

4.5 Заключение

В тази глава от дисертацията разгледахме корелиран "probit" модел за анализиране на няколко наредени категорни променливи. Описахме разширение на ECM алгоритъма от главата Модел за една наредена категорна променлива с повторни наблюдения 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 5

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: Модел за една наредена категорна променлива с повторни наблюдения и Съвместен модел за няколко наредени категорни променливи. Резултатите, изложени тук, не са публикувани до момента.

5.1 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:

$$\begin{aligned} y_{1ij} &= \boldsymbol{x'_{1ij}\beta_1} + \boldsymbol{z'_{1ij}b_{1i}} + \epsilon_{1ij}, \\ y_{2ij} &= \boldsymbol{x'_{2ij}\beta_2} + \boldsymbol{z'_{2ij}b_{2i}} + \epsilon_{2ij}. \end{aligned} \tag{5.1}$$

Връзката между скритата нормална и наблюдаваната категорна променлива е следната:

$$y^*_{1ij} = \begin{cases} 1, \ y_{1ij} \leq \alpha_1; \\ l, \ \alpha_{l-1} < y_{1ij} \leq \alpha_l, \ l = 2, \ldots, m-1; \\ m, \ y_{1ij} > \alpha_{m-1}; \end{cases} \tag{5.2}$$

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 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 5.1 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 $\boldsymbol{z_{2ij}}$. Ковариационната матрица на случайните ефекти $\boldsymbol{\Sigma}$ е квадратна матрица с размерност $q \times q$ и

е положително полу-дефинитна.

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 нееднозначна 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. Възможни са други ограничения за параметрите на модела, които да доведат до пълна определеност.

5.2 МПО чрез EM алгоритъм

Идеята на подхода за намиране на МПО на неизвестните параметри в модела 5.1 и неизвестните прагове в 5.2 чрез EM алгоритъм е същата като тази, описана в предходните Глава 3 и Глава 4. За начало въвеждаме в явен вид праговете във функцията на пълното правдоподобие. За тази цел използваме подхода на Kawakatsu and Largey [2009], които развиват работата на Ruud [1991]. 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.

5.2.1 Лог-правдоподобие на пълния набор от данни

Пълното лог-правдоподобие $\ln L$ е многомерната съвместна плътност на случайните ефекти, наблюдаваната непрекъсната величина и трансформираната скрита величина:

$$\ln L \quad = \quad \ln f(\boldsymbol{b}, \boldsymbol{y_{1_{new}}}, \boldsymbol{y_2}) = \sum_{i=1}^{n} \ln f(\boldsymbol{b_i}) f(\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}} | \boldsymbol{b_i})$$

$$= \sum_{i=1}^{n} \ln[f(\boldsymbol{b_i}) \prod_{j=1}^{n_i} f(y_{2ij}|\boldsymbol{b_i}) f(y_{1ij_{new}}|\boldsymbol{b_i}, y_{2ij})],$$

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:

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

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.

5.2.2 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:

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

Следва, че регресионните параметри $\boldsymbol{\beta_1}$ са решение по МНМК при регресия на $\tilde{y}_{1ij}$ по $\boldsymbol{x_{1ij}}$, където $\tilde{y}_{1ij} = \delta_{y^*_{1ij}} y_{1ij_{new}} - \boldsymbol{z'_{1ij} b_{1i}} + \alpha_{y^*_{1ij}-1} - \lambda(y_{2ij} - \mu_{2ij})$.

Регресионните параметри за фиксираните ефекти $\boldsymbol{\beta_2}$ за наблюдаваната нормална променлива удовлетворяват следната система от уравнения:

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

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:

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

Следва, че $\hat{\lambda}$ е решение по МНМК при регресия на $(\delta_{y^*_{1ij}} y_{1ij_{new}} - \delta_{y^*_{1ij}} \mu^*_{1ij})$ по $(y_{2ij} - \mu_{2ij})$.

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. Доказателството следва в следващия параграф.

5.2.3 Условни очаквания

Нека използваме следните означения за опростяване на записа:

$$\boldsymbol{X_{1i}} = \begin{pmatrix} \boldsymbol{x'_{1i1}} \\ \boldsymbol{x'_{1i2}} \\ \vdots \\ \boldsymbol{x'_{1in_i}} \end{pmatrix}, \boldsymbol{X_{2i}} = \begin{pmatrix} \boldsymbol{x'_{2i1}} \\ \boldsymbol{x'_{2i2}} \\ \vdots \\ \boldsymbol{x'_{2in_i}} \end{pmatrix}, \boldsymbol{Z_{1i}} = \begin{pmatrix} \boldsymbol{z'_{1i1}} \\ \boldsymbol{z'_{1i2}} \\ \vdots \\ \boldsymbol{z'_{1in_i}} \end{pmatrix}, \boldsymbol{Z_{2i}} = \begin{pmatrix} \boldsymbol{z'_{2i1}} \\ \boldsymbol{z'_{2i2}} \\ \vdots \\ \boldsymbol{z'_{2in_i}} \end{pmatrix},$$

$$\boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{Z_{1i}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z_{2i}} \end{pmatrix}, \boldsymbol{\alpha_i} = \begin{pmatrix} \alpha_{y^*_{1i1}-1} \\ \alpha_{y^*_{1i2}-1} \\ \vdots \\ \alpha_{y^*_{1in_i}-1} \end{pmatrix}, \boldsymbol{y_{1i_{new}}} = \begin{pmatrix} y_{1i1_{new}} \\ y_{1i2_{new}} \\ \vdots \\ y_{1in_{i_{new}}} \end{pmatrix}, \boldsymbol{y_{2i}} = \begin{pmatrix} y_{2i1} \\ y_{2i2} \\ \vdots \\ y_{2in_i} \end{pmatrix},$$

$$\boldsymbol{\delta_i^{-1}} = \begin{pmatrix} 1/\delta_{y^*_{1i1}} \\ 1/\delta^*_{y_{1i2}} \\ \vdots \\ 1/\delta_{y^*_{1in_i}} \end{pmatrix}.$$

Тогава съвместното разпределение на $(\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}})$ и $\boldsymbol{b_i}$ е многомерно нормално:

$$\begin{pmatrix} \boldsymbol{y_{1i_{new}}} \\ \boldsymbol{y_{2i}} \\ \boldsymbol{b_i} \end{pmatrix} \sim N \left[\begin{pmatrix} (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} \\ \boldsymbol{X_{2i}\beta_2} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{B'} & \boldsymbol{D} & \boldsymbol{E} \\ \boldsymbol{C'} & \boldsymbol{E'} & \boldsymbol{\Sigma} \end{pmatrix} \right],$$

където

$$\begin{aligned}
\boldsymbol{A} &= (\boldsymbol{Z_{1i}\Sigma_{11}Z'_{1i}} + \boldsymbol{I_{n_i}} \otimes \boldsymbol{\sigma_{11}}) \circ \boldsymbol{\delta_i^{-1}\delta_i^{-1'}},\\
\boldsymbol{B} &= (\boldsymbol{Z_{1i}\Sigma_{12}Z'_{2i}} + \boldsymbol{I_{n_i}} \otimes \boldsymbol{\sigma_{12}}) \circ (\boldsymbol{J_{n_i\times q}\delta_i^{-1'}}),\\
\boldsymbol{C} &= \boldsymbol{Z_{1i}}(\boldsymbol{\Sigma_{11}}\ \boldsymbol{\Sigma_{12}}) \circ (\boldsymbol{J_{n_i\times q}\delta_i^{-1}}),\\
\boldsymbol{D} &= (\boldsymbol{Z_{2i}\Sigma_{22}Z'_{2i}} + \boldsymbol{I_{n_i}} \otimes \boldsymbol{\sigma_{22}}),\\
\boldsymbol{E} &= \boldsymbol{Z_{2i}}(\boldsymbol{\Sigma_{21}}\ \boldsymbol{\Sigma_{22}}),
\end{aligned}$$

$\boldsymbol{J_{n_i\times q}\delta_i^{-1}}$ е $n_i \times q$ матрица с колони $\boldsymbol{\delta_i^{-1}}$, $\boldsymbol{I_{n_i}}$ е n_i-мерна единична матрица, $\circ$ е поелементно умножение (умножение на Адамар) и $\otimes$ е умножение на Кронекер.

Нека да означим с $\boldsymbol{M_i} = \begin{pmatrix} \boldsymbol{y_{1i_{new}}} - (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} \\ \boldsymbol{y_{2i}} - \boldsymbol{X_{2i}\beta_2} \end{pmatrix}$ и със

$\boldsymbol{\Sigma_{B_i}} = \begin{pmatrix} \boldsymbol{C'} & \boldsymbol{E'} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B'} & \boldsymbol{D} \end{pmatrix}^{-1}$.

Тогава условното разпределение на $\boldsymbol{b_i}$ при условие $(\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}})$ е отново нормално:

$$\boldsymbol{b_i}|(\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}}) \sim N[\boldsymbol{\Sigma_{B_i}M_i}, \boldsymbol{\Sigma} - \boldsymbol{\Sigma_{B_i}} \begin{pmatrix} \boldsymbol{C} \\ \boldsymbol{E} \end{pmatrix}] .$$

В изразите на оценките на Е-стъпката на алгоритъма участват следните условни очаквания: $E(\boldsymbol{b_i}|\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}}), E(\boldsymbol{b_i b'_i}|\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}}), E(y_{1ij_{new}}\boldsymbol{b_i}|\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}})$. Ще покажем, че те зависят само от първите два момента на $\boldsymbol{y_{1i_{new}}}|(\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}})$, което е орязано нормално разпределение.

Очакването на случайните ефекти при условие наблюдаваните данни е:

$$\begin{aligned}
E(\boldsymbol{b_i}|(\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}})) &= E[E(\boldsymbol{b_i}|(\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}}))|(\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}})]\\
&= E[\boldsymbol{\Sigma_{B_i}} \begin{pmatrix} \boldsymbol{y_{1i_{new}}} - (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} \\ \boldsymbol{y_{2i}} - \boldsymbol{X_{2i}\beta_2} \end{pmatrix} |(\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}})]\\
&= \boldsymbol{\Sigma_{B_i}} \begin{pmatrix} E(\boldsymbol{y_{1i_{new}}}|(\boldsymbol{y^*_{1i}}, \boldsymbol{y_{2i}})) - (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} \\ \boldsymbol{y_{2i}} - \boldsymbol{X_{2i}\beta_2} \end{pmatrix}.
\end{aligned}$$

Очакването на втория момент на случайните ефекти при условие наблю-

даваните данни е:

$$\begin{aligned} E(\boldsymbol{b_i b_i'}|(\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}})) &= E[E(\boldsymbol{b_i b_i'}|\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}})|(\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}})] \\ &= E[Var(\boldsymbol{b_i}|\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}}) + E(\boldsymbol{b_i}|\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}})E(\boldsymbol{b_i'}|\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}})|\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}}]. \end{aligned}$$

В израза за новите оценки на праговете участва следното условно очакване:

$$\begin{aligned} E(y_{1ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}}) &= E[E(y_{1ij_{new}}\boldsymbol{b_i}|\boldsymbol{y_{1i_{new}}}, \boldsymbol{y_{2i}})|\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}}] \\ &= E(y_{1ij_{new}}\boldsymbol{\Sigma_{B_i}} \begin{pmatrix} \boldsymbol{y_{1i_{new}}} - (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} \\ \boldsymbol{y_{2i}} - \boldsymbol{X_{2i}\beta_2} \end{pmatrix} |\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}}) \\ &= \boldsymbol{\Sigma_{B_i}} \begin{pmatrix} E(y_{1ij_{new}}^2|\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}}) - (\boldsymbol{X_{1i}\beta_1} - \boldsymbol{\alpha_i}) \circ \boldsymbol{\delta_i^{-1}} \\ E(y_{1ij_{new}}|\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}})(\boldsymbol{y_{2i}} - \boldsymbol{X_{2i}\beta_2}) \end{pmatrix}. \end{aligned}$$

5.2.4 $(k+1)$-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 $E(\tilde{y}_{1ij}|\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}}; \boldsymbol{\Gamma^k})$ по $\boldsymbol{x_{1ij}}$.

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

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

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

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

  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: $\delta_u^{k+1} =$

$$\frac{-E[b|\boldsymbol{y_1^*}, \boldsymbol{y_2}; \boldsymbol{\Gamma^k}] + \sqrt{(E[b|\boldsymbol{y_1^*}, \boldsymbol{y_2}; \boldsymbol{\Gamma^k}]^2 - 4E[a|\boldsymbol{y_1^*}, \boldsymbol{y_2}; \boldsymbol{\Gamma^k}]E[c|\boldsymbol{y_1^*}, \boldsymbol{y_2}; \boldsymbol{\Gamma^k}])}}{2E[a|\boldsymbol{y_1^*}, \boldsymbol{y_2}; \boldsymbol{\Gamma^k}]}.$$

  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 уравнения:

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

  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 $\hat{\boldsymbol{\Sigma}}^{k+1} = \frac{1}{n}\sum_{i=0}^{n} E(\boldsymbol{b_i}\boldsymbol{b'_i}|\boldsymbol{y_{1i}^*}, \boldsymbol{y_{2i}}; \boldsymbol{\Gamma^k})$. 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.

5.2.5 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. Стъпките са следните:

1. Оценяваме модела 5.1 и неизвестните прагове в 5.2 според данните, състоящи се от n обекта, използвайки предложения ECM алгоритъм и получаваме оценки за неизвестните параметри, означени с $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\delta}}, \hat{\lambda}, \hat{\sigma_{22}})$. За да генерираме "bootstrap" извадка, първо генерираме n случайни ефекта $\boldsymbol{b}_{\boldsymbol{k}}^{\boldsymbol{b}}$ от $N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, $k = 1, \ldots, n$. След това симулираме нормални вектори $\boldsymbol{y}_{\boldsymbol{k}}^{\boldsymbol{b}}$ с размерност n_k според модел 5.1 за всеки случаен ефект $\boldsymbol{b}_{\boldsymbol{k}}^{\boldsymbol{b}}$. Използваме оценените прагове $\hat{\boldsymbol{\delta}}$, за да определим в кой интервал попадат нормалните данни $\boldsymbol{y}_{\mathbf{1}\boldsymbol{k}}^{\boldsymbol{b}}$, $k = 1, \ldots, n$ и използвайки 5.2 определяме нивото на "bootstrap" категорната променлива $\boldsymbol{y}_{\mathbf{1}\boldsymbol{k}}^{\boldsymbol{b}*}$. Така получаваме "bootstrap" извадка, състояща се от категорните променливи $\boldsymbol{y}_{\mathbf{1}\boldsymbol{k}}^{\boldsymbol{b}*}$, $k = 1, \ldots, n$ и нормалните променливи $\boldsymbol{y}_{\mathbf{2}\boldsymbol{k}}^{\boldsymbol{b}}$, $k = 1, \ldots, n$.

2. Прилагаме EM алгоритъма към получените "bootstrap" данни $\boldsymbol{y}_{\mathbf{1}\boldsymbol{k}}^{\boldsymbol{b}*}, \boldsymbol{y}_{\mathbf{2}\boldsymbol{k}}^{\boldsymbol{b}}$, $k = 1, \ldots, n$, за да получим оценки $\boldsymbol{\Gamma}^{\boldsymbol{b}}$ за генерираното множество от данни.

3. Използваме Монте Карло метод, за да намерим приближение на "bootstrap" ковариационната матрица. Повтаряме стъпка 1 и стъпка 2 B пъти и пресмятаме ковариационната матрица на B оценени параметри $\boldsymbol{\Gamma}^{\boldsymbol{b}}, b = 1, \ldots, B$:
$$Cov(\hat{\boldsymbol{\Gamma}}) \approx \sum_{b=1}^{B} \frac{(\boldsymbol{\Gamma}^{\boldsymbol{b}} - \bar{\boldsymbol{\Gamma}})(\boldsymbol{\Gamma}^{\boldsymbol{b}} - \bar{\boldsymbol{\Gamma}})'}{B-1},$$
където $\bar{\boldsymbol{\Gamma}} = \sum_{b=1}^{B} \boldsymbol{\Gamma}^{\boldsymbol{b}}/B$.

5.3 Симулации

За реализирането на алгоритъма сме използвали безплатната софтуерна среда за статистически изчисления и графики **R**. Кодът на **R** е представен в Приложение B: R код за ECM алгоритъм за корелиран "probit" модел за една наредена категорна и една нормална величини от дългосрочни проучвания.

Симулирани са стойности от следния модел със случайни свободни членове:

$$y_{1ij} \;=\; \beta_{10} + \beta_{11} t_{ij} + b_{1i} + \epsilon_{1ij},\; j = 1, \ldots, 4,$$

$$y_{2ij} \quad = \quad \beta_{20} + \beta_{21} t_{ij} + b_{2i} + \epsilon_{2ij}, \;\; j = 1, \ldots, 4, \tag{5.3}$$

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:

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

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, което е в рамките на препоръчителния брой между 50 и 100 “bootstrap” повторения (Efron and Tibshirani [1994]). Резултатите са представени в Таблица 5.1.

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.

Таблица 5.1: Таблица с оценките и стандартните грешки от двете симулационни изследвания за "probit" модела 5.3 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

5.4 Приложение на модела

Прилагаме предложения модел 5.1 към данните от ЗПП (използвани в секция Приложение на модела на Глава 3). 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:

$$
\begin{aligned}
y_{ij} &= \beta_{10} + \beta_{11} t_{ij} + b_{1i} + \epsilon_{1ij}, j = 1, \ldots, 7 \\
\mathrm{bmi}_{ij} &= \beta_{20} + \beta_{21} t_{ij} + b_{2i} + \epsilon_{2ij}, j = 1, \ldots, 7 \\
\mathrm{srh}_{ij} &= \begin{cases} 1, \ y_{ij} \leq \alpha_1 = 0, \\ j, \ \alpha_{j-1} < y_{ij} \leq \alpha_j, \ j = 2, \ldots, m-1, \\ 5, \ y_{ij} > \alpha_{m-1}, \end{cases}
\end{aligned}
\tag{5.4}
$$

където ковариационната матрица на грешките е:

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

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 5.2.[1]

Таблица 5.2: Таблица с оценки и стандартни грешки от съвместния корелиран “probit” модел 5.4 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 5.2 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.

Таблица 5.3: Таблица с оценки и стандартни грешки от корелирания "probit" модел 5.5 за самооценка за здравето, приложен към данните от здравно и пенсионно проучване.

параметри	β_{10}	β_{11}	δ_2	δ_3	δ_4	σ^b_{11}
оценки	1.203	0.115	1.576	1.492	1.414	2.147
стд. гр.	0.0159	0.0026	0.0095	0.0100	0.0130	0.0467
z-score	75.70	44.54	166.01	148.83	108.70	45.94

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. Разликите в индивидите относно индекса на телесна маса е значително по-голяма, отколкото за самооценката за здравето.

Бихме могли (с помощта на ЕМ алгоритъма от Глава 3) да оценим следния самостоятелен модел за самооценка за здравето:

$$\begin{aligned} y_{ij} &= \beta_{10} + \beta_{11} t_{ij} + b_i + \epsilon_{ij}, \\ \text{srh}_{ij} &= \begin{cases} 1, \ y_{ij} \leq \alpha_1 = 0, \\ l, \ \alpha_{l-1} < y_{ij} \leq \alpha_l, \ l = 2, 3, 4, \\ 5, \ y_{ij} > \alpha_4, \end{cases} \end{aligned} \tag{5.5}$$

където $\epsilon_{ij} \sim N(0,1), j = 1, \ldots, 7$, $Var(b_i) = \sigma^b_{11}$ и с 'srh_{ij}' сме отбелязали самооценка за здравето на $i-$ти индивид в момент от време j. Резултатите са представени в Таблица 5.3.

Самостоятелният модел за индекса на телесна маса е следният:

$$\text{bmi}_{ij} = \beta_{20} + \beta_{21} t_{ij} + b_i + \epsilon_{ij}, j = 1, \ldots, 7, \tag{5.6}$$

където $\epsilon_{ij} \sim N(0, \sigma_{22}), j = 1, \ldots, 7$, $Var(b_i) = \sigma^b_{22}$ и с 'bmi_{ij}' сме отбелязали индекса на телесна маса на $i-$ти индивид в момент от време j. Резултатите са представени в Таблица 5.4.

От Таблици 5.2, 5.3 и 5.4 се вижда, че оценките и стандартните грешки за параметрите в съвместния и отделните модели са много близки по стойност. Забелязва се, че в отделния модел оценките на параметрите за самооцен-

Таблица 5.4: Таблица с оценки и стандартни грешки от корелирания "probit" модел 5.6 за индекса на телесна маса, приложен към данните от здравно и пенсионно проучване.

параметри	β_{20}	β_{21}	σ_{22}	σ^b_{22}
оценки	26.971	0.149	3.276	23.661
стд. гр.	0.0590	0.0043	0.0223	0.4367
z-score	457.44	34.79	146.60	54.18

ката за здраве са по-ефективни от тези в съвместния модел с изключение на оценката за дисперсията на случайния свободен член. Докато оценките на параметрите за индекса на телесна маса са по-ефективни в съвместния модел сравнени с отделния (изключение е дисперсията на грешката). Разликите са малки, което се дължи най-вероятно на големия брой индивиди и наблюдения над обект.

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. Gueorguieva and Sanacora [2006] показват чрез симулационно изследване, че оценките на параметрите за категорната величина в съвместния модел са по-малко изместени от оценките на параметрите за категорната величина от отделния модел. В същата статия на база друго симулационно изследване с липсващи наблюдения за непрекъснатата величина е показано, че оценките на параметрите за непрекъсната величина в съвместния модел са малко по-ефективни от тези в отделния модел.

5.5 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 6

Заключение

6.1 Научен и практически принос на дисертацията

В настоящата дисертация 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.

6.2 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. Проблемът е обобщен в статията на Hogan et al. [2004]. В нея основно са разгледани три типа модели за съвместното разпределение на променливата от основен интерес и механизма на липсите условно по предикторите. Като първи е изложен селективният модел, разгледан от Diggle and Kenward [1994]. На второ място е представен смесеният модел на Little and Wang [1996]. Последен е моделът със случаен член, който използва скрита променлива за създаване на връзка между величината от основен интерес и индикаторна променлива за отпадане на обекта. Ранен пример за такъв модел е този, разгледан от Wu and Caroll [1987]. Статията на Hogan et al. [2004] дава референции към други методи (полу-параметрични и непараметрични) за взимане предвид на липсата на наблюдения.

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 полупараметричният модел на Кокс с пропорционални рискове (Therneau and Grambsch [2001]). Статията на Guo and Carlin [2004] разглежда подхода на Henderson et al. [2000] и го сравнява със стандартния подход за отделното моделиране на двете променливи. В статистиката с медицинска насоченост най-често се моделира времето до смърт, излекуване от дадена болест, ремисия, достигане на дадени нива на тип клетки в кръвта. Моделирането на времето до настъпване на някакво събитие е дял от статистиката, наречен анализ на оцеляването. Най-популярният модел за времето до настъпване на събитие е попу-параметричният модел на Кокс, но са възможни и чисто параметрични (пример са експоненциален модел, модел на Вайбул) или чисто непараметрични подходи. 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.

В литературата има разгледани съвместни модели на променлива, мерена във времето и време до настъпване на няколко различни събития (Elashoff et al. [2007, 2008]). В статиите на McGilchrist [1993] и McGilchrist and Aisbett [1991] са описани модели на Кокс със случайни ефекти и начините за оценяването им. Други статии, където е разгледан този проблем, са статиите на Klein [1992], Nielsen et al. [1992], Verweij and Van Houwelingen [1994], Ripatti and Palmgren [2000].

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].

Такъв един модел би бил от полза предвид нарастващата нужда от съвместното моделиране на няколко променливи, проследявани във времето. Отпадането на обектите от изследването е почти неизбежно усложнение при този тип изследвания. Коректният анализ налага да се вземе предвид този факт. Това налага доразвиване на вече съществуващите методи. Използването на безплатна среда за разработване ще даде предпоставка за по-широка употреба, по-бързо развитие и подобряване на модела.

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.

Приложение А: R код за ECM алгоритъм за корелиран “probit” модел за една наредена категорна величина от дългосрочни проучвания

```
##################################################
######################   ecm algorithm
#########################################################
ecm.rand.int=function(start.values.beta,start.values.delta,start.values.sigma.rand,
montecarlo,data.ordinal,predictors.fixed,predictors.random,epsilon) {

#########################################
### wide format of longitudinal data
### predictors.fixed is a list
### predictors.random is an array (matrix)
### (first columns are random predictors for the first consecutive observation,
### then for the second and so on)
### sigma.rand is the covaraince matrix of the random effects
##################################

library(MASS)
library(tmvtnorm)
```

```
betanew=start.values.beta
betaold=start.values.beta-2
deltanew=start.values.delta
deltaold=start.values.delta-2
sigma.rand.new=start.values.sigma.rand
sigma.rand.old=start.values.sigma.rand+0.1
number.it=0

# definition of lower and upper boundary of the truncated normal distr.
# (the new variable given observed data)
trunc.lower=ifelse(data.ordinal==1,-Inf,0)
trunc.upper=ifelse(data.ordinal==1,0,ifelse(data.ordinal<5,1,Inf))
#number of observations in each category
n=0
for (i in 1:5) n[i]=sum(data.ordinal==i)
observations=length(data.ordinal[,1])
mult.obs=length(predictors.fixed)
dimension.sigma=ifelse(length(sigma.rand.old)==1,1,ncol(sigma.rand.old))
#################################################################
############             while loop           ##########################
#######################################################################

while(sum(abs(c(betanew-betaold,deltanew-deltaold,sigma.rand.new-sigma.rand.old))
>epsilon)>0) {

number.it=number.it+1
betaold=betanew
deltaold=deltanew
sigma.rand.old=sigma.rand.new

#######################################################################

delta=ifelse(data.ordinal==1 |data.ordinal==5,1,
             ifelse(data.ordinal==2, deltaold[1],
```

```
                    ifelse(data.ordinal==3, deltaold[2], deltaold[3])
                    )
               )
alpha=ifelse(data.ordinal==1 |data.ordinal==2,0,
             ifelse(data.ordinal==3, deltaold[1],
                    ifelse(data.ordinal==4, sum(deltaold[1:2]), sum(deltaold))
                    )
               )

###expectation of ynew
pred=sapply(1:mult.obs, function(i) (predictors.fixed[[i]]%*%betaold), simplify=TRUE)
## makes pred a mtarix: observations x mult.obs
mean.ynew=(pred-alpha)/delta

if(dimension.sigma==1) sigma.rand.old=matrix(sigma.rand.old,ncol=dimension.sigma)
## makes sigma.rand.old a matrix in the case when we have only random intercept model

z=sapply(1:observations, function(i) matrix(predictors.random[i,],
ncol=dimension.sigma,byrow=T), simplify="array")
variance.ynew=sapply(1:observations, function(i)
(z[,,i]%*%sigma.rand.old%*%t(z[,,i])+diag(mult.obs))/(delta[i,]%*%t(delta[i,])),
simplify="array")

### simulations
simulations=sapply(1:observations, function(i) rtmvnorm(montecarlo,
mean=mean.ynew[i,], sigma=matrix(as.vector(variance.ynew[,,i]),ncol=mult.obs),
lower=trunc.lower[i,],upper=trunc.upper[i,], algorithm="gibbs"),
simplify="array")
## montecarlo x mult.obs x observations
moments1=t(sapply(1:observations, function(i) apply(simulations[,,i],2,mean),
simplify="array"))
moments2=sapply(1:observations, function(i) var(simulations[,,i]),simplify="array")

######### to calculate expectation of random effects
sigmab=sapply(1:observations, function(i) t(t(sigma.rand.old%*%t(z[,,i]))/delta[i,]
```

```
)%*%solve(variance.ynew[,,i]),simplify="array")
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%((moments1[i,])-
mean.ynew[i,]), simplify="array")

if (dimension.sigma>1)
{ywave=sapply(1:observations, function(i) delta[i,]*(moments1[i,])-z[,,i]%*%
firstmomentb[,,i]+alpha[i,],simplify=F) }      else {ywave=sapply(1:observations,
function(i) delta[i,]*(moments1[i,])-
z[,,i]%*%t(firstmomentb[i])+alpha[i,], simplify=F) }
#ywave is a list of vector column for each object in the study

betanew=lm(c(t(do.call(cbind,ywave)))~do.call(rbind,predictors.fixed)-1)$coef
print(betanew)
################################
#### update
############################
pred=sapply(1:mult.obs, function(i) predictors.fixed[[i]]%*%betanew, simplify=TRUE)
mean.ynew=(pred-alpha)/delta
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%((moments1[i,])-
mean.ynew[i,]), simplify="array")

#######################################
################### delta estimates
##########################################

first=t(sapply(1:observations, function(i) diag(moments2[,,i])+(moments1[i,])^2,
simplify="array"))

if (dimension.sigma>1)
{pred.random=t(sapply(1:observations, function(i) z[,,i]%*%firstmomentb[,,i],
simplify=T)) }      else {pred.random=t(sapply(1:observations, function(i)
z[,,i]%*%t(firstmomentb[i]),simplify=T)) }

second1=moments1*pred
```

```
second2=sapply(1:observations, function(i)
sigmab[,,i]%*%(moments2[,,i]+moments1[i,]%*%t(moments1[i,])-mean.ynew[i,]%*%
t(moments1[i,])),
simplify="array")
## for each i I have a q x n_i matrix - I need a column for each j
#### I need the columns in the last part to be y_ijnew*mean.ynew

if (dimension.sigma==1) second2=t(sapply(1:observations, function(i) diag(
z[,,i]%*%t(second2[,,i])),simplify="array")) else second2=t(sapply(1:
observations, function(i) diag(z[,,i]%*%second2[,,i]),simplify="array"))

for (k in 1:3) {

deltaold=deltanew
delta=ifelse(data.ordinal==1 |data.ordinal==5,1,
             ifelse(data.ordinal==2, deltaold[1],
                   ifelse(data.ordinal==3, deltaold[2], deltaold[3])
                   )
              )

alpha=ifelse(data.ordinal==1 |data.ordinal==2,0,
             ifelse(data.ordinal==3, deltaold[1],
                   ifelse(data.ordinal==4, deltaold[1]+deltaold[2], sum(deltaold))
                   )
              )

mean.ynew=(pred-alpha)/delta
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%((moments1[i,])-
mean.ynew[i,]), simplify="array")

third=moments1*delta-pred-pred.random
second3=-moments1*alpha

a=sum(first[data.ordinal==(k+1)])+sum(n[(k+2):5])
b=0
```

```
if(k==1) b=sum((second1+second2+second3)[data.ordinal==2])
-sum(third[data.ordinal==3])-sum(third[data.ordinal==4]+deltaold[2])-
sum(third[data.ordinal==5]+sum(deltaold[2:3])) else if (k==2) b=sum((second1+second2+
second3)[data.ordinal==3])-sum(third[data.ordinal==4]+deltanew[1])-
sum(third[data.ordinal==5]+deltanew[1]+deltaold[3]) else b=sum((second1+
second2+second3)[data.ordinal==4])-sum(third[data.ordinal==5]+sum(deltanew[1:2]))

c=-n[(k+1)]

deltanew[k]=(b+sqrt(b^2-4*a*c))/(2*a)
print(deltanew[k])
}
###########################
################ update
############################
deltaold=deltanew
delta=ifelse(data.ordinal==1 |data.ordinal==5,1,
             ifelse(data.ordinal==2, deltaold[1],
                    ifelse(data.ordinal==3, deltaold[2], deltaold[3])
                    )
              )
alpha=ifelse(data.ordinal==1 |data.ordinal==2,0,
             ifelse(data.ordinal==3, deltaold[1],
                    ifelse(data.ordinal==4, deltaold[1]+deltaold[2], sum(deltaold))
                    )
              )
mean.ynew=(pred-alpha)/delta
sigmab=sapply(1:observations, function(i) t(t(sigma.rand.old%*%t(z[,,i]))/
delta[i,])%*%solve(variance.ynew[,,i]),simplify="array")
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%((moments1[i,])-
mean.ynew[i,]), simplify="array")

############################### sigma.rand estimate
if(dimension.sigma==1)
sigma.rand=sapply(1:observations, function(i)
```

```
sigma.rand.old-sigmab[,,i]%*%((z[,,i]%*%sigma.rand.old)/delta[i,])+
sigmab[,,i]%*%(moments2[,,i]+(moments1[i,])%*%t(moments1[i,])-
(moments1[i,])%*%t(mean.ynew[i,])
-(mean.ynew[i,])%*%t(moments1[i,])+(mean.ynew[i,])%*%t(mean.ynew[i,])
)%*%sigmab[,,i], simplify=T) else
sigma.rand=sapply(1:observations, function(i)
sigma.rand.old-sigmab[,,i]%*%((z[,,i]%*%sigma.rand.old)/delta[i,])+
sigmab[,,i]%*%(moments2[,,i]+(moments1[i,])%*%t(moments1[i,])-
(moments1[i,])%*%t(mean.ynew[i,])
-(mean.ynew[i,])%*%t(moments1[i,])+(mean.ynew[i,])%*%t(mean.ynew[i,])
)%*%t(sigmab[,,i]), simplify=T)

if(dimension.sigma==1) sigma.rand.new=mean(sigma.rand) else sigma.rand.new=
matrix(apply(sigma.rand,1,mean),ncol=dimension.sigma)

print(sigma.rand.new)

#######################
print(c(number.it,sigma.rand.new,betanew,deltanew))
} # while

c(sigma.rand.new,betanew,deltanew)
 } #function


##############################################################################
################### data simulation
##############################################################################
##################  RI model
#################################################
random.int=rnorm(500,0,0.1)
l=length(random.int)
int=-0.5
y1=sapply(1:5,function(i) random.int+int+i+rnorm(l), simplify="array")
data.ordinal=ifelse(y1<=0,1,ifelse(y1<=1.5,2,ifelse(y1<=3,3,ifelse(y1<=4,4,5))))
```

```
#data.ordinal
table(data.ordinal)
time=sapply(1:5, function(i) rep(i,l), simplify="array")
predictors=sapply(1:5, function(i) cbind(1,time[,i]), simplify="array")
predictors.fixed=sapply(1:5, function(i) predictors[,,i], simplify=F)
predictors.random=matrix(rep(time[,1],5),ncol=5)

start.values.beta=c(-0.55,0.95)
start.values.delta=c(1.5,1.5,1)
start.values.sigma.rand=.01
montecarlo=200

example1=ecm.rand.int(start.values.beta,start.values.delta,start.values.sigma.rand,
montecarlo=1500,data.ordinal,predictors.fixed,predictors.random,epsilon=0.0005)


##########################
########## RIAS model
#########################################
random.effects=mvrnorm(n=5000, mu=c(0,0),Sigma=matrix(c(0.0001,0.00027,0.00027,
0.0009),ncol=2))
l=length(random.effects)/ncol(random.effects)
time=1:15
beta=c(0.5,1,2,-2.5)
pred1=rnorm(l)
pred2=rexp(l)
predictors.fixed=sapply(1:15, function(i) cbind(1,rep(i,l),pred1[i],pred2[i]),
 simplify=F)
y1=t(sapply(1:l, function(j) sapply(1:15, function(i) c(1,time[i])%*%
random.effects[i,]+predictors.fixed[[i]][j,]%*%beta+rnorm(1)), simplify="array"))
data.ordinal=ifelse(y1<=0,1,ifelse(y1<=1.5,2,ifelse(y1<=3,3,ifelse(y1<=4,4,5))))
table(data.ordinal)
predictors.random=do.call(cbind,sapply(1:15,function(i) cbind(1,rep(i,l)),simplify=F))

start.values.beta=c(0.5,1,2,-2.5)
```

```
start.values.delta=c(1.5,1.5,1)
start.values.sigma.rand=matrix(c(0.0001,0.00027,0.00027,0.0009),ncol=2)

ecm.rand.int(start.values.beta,start.values.delta,start.values.sigma.rand,
montecarlo=150,data.ordinal,predictors.fixed,predictors.random,epsilon=0.0006)->fff
```

Приложение Б: R код за ECM алгоритъм за корелиран “probit” модел за две наредени категорни величини

```
##################################################
######################   ecm algorithm
#######################################################
ecm.mult.ord=function(start.values.beta1,start.values.beta2,
start.values.delta1,start.values.delta2,start.values.sigma.rand,
data.ordinal1,data.ordinal2,predictors.fixed1,predictors.fixed2,
predictors.random1,predictors.random2,epsilon) {

#########################################
### sigma.rand is the covaraince matrix of the random effects
##################################
library(MASS)
library(tmvtnorm)

betanew1=start.values.beta1
betaold1=start.values.beta1-2
betanew2=start.values.beta2
betaold2=start.values.beta2-2
```

```
deltanew1=start.values.delta1
deltaold1=start.values.delta1-2
deltanew2=start.values.delta2
deltaold2=start.values.delta2-2

sigma.rand.new=start.values.sigma.rand
sigma.rand.old=start.values.sigma.rand+0.01

number.it=0

# definition of lower and upper boundary of the truncated normal distr.
#(the new variable given observed data)
data.ordinal=cbind(data.ordinal1,data.ordinal2)
trunc.lower=ifelse(data.ordinal==1,-Inf,0)
trunc.upper=ifelse(data.ordinal==1,0,ifelse(data.ordinal<3,1,Inf))
#number of observations in each category
n1=0
for (i in 1:3) n1[i]=sum(data.ordinal1==i)
n2=0
for (i in 1:3) n2[i]=sum(data.ordinal2==i)

observations=length(data.ordinal1)
mult.obs=2
dimension.sigma=ifelse(length(sigma.rand.old)==1,1,ncol(sigma.rand.old))

#################################################################
############             while loop           ##########################
#######################################################################

while(sum(abs(c(betanew1-betaold1,deltanew1-deltaold1,betanew2-betaold2,
deltanew2-deltaold2,sigma.rand.new-sigma.rand.old))>epsilon)>0) {

number.it=number.it+1
betaold1=betanew1
deltaold1=deltanew1
```

```
betaold2=betanew2
deltaold2=deltanew2
sigma.rand.old=sigma.rand.new
#######################################################################

delta1=ifelse(data.ordinal1==1 |data.ordinal1==3,1,deltaold1)
delta2=ifelse(data.ordinal2==1 |data.ordinal2==3,1,deltaold2)
delta=cbind(delta1,delta2)

alpha1=ifelse(data.ordinal1==1 |data.ordinal1==2,0, deltaold1)
alpha2=ifelse(data.ordinal2==1 |data.ordinal2==2,0, deltaold2)
alpha=cbind(alpha1,alpha2)

###expectation of ynew
predynew1=predictors.fixed1%*%betaold1
mean.ynew1=(predynew1-alpha1)/delta1

if(length(betaold2)>1) predynew2=predictors.fixed2%*%betaold2 else predynew2=
predictors.fixed2*betaold2
mean.ynew2=(predynew2-alpha2)/delta2
mean.ynew=cbind(mean.ynew1,mean.ynew2)

if(dimension.sigma==1) sigma.rand.old=matrix(sigma.rand.old,ncol=dimension.sigma)
## makes sigma.rand.old a matrix in the case when we have only random intercept model

predictors.random=cbind(predictors.random1,predictors.random2)
z=sapply(1:observations, function(i) diag(predictors.random[i,]), simplify="array")
variance.ynew=sapply(1:observations, function(i)
(z[,,i]%*%sigma.rand.old%*%t(z[,,i])+diag(mult.obs))/(delta[i,]%*%t(delta[i,])),
simplify="array")

### computation of exact moments
simulations=sapply(1:observations, function(i) mtmvnorm(mean=mean.ynew[i,],
sigma=matrix(as.vector(variance.ynew[,,i]),ncol=mult.obs),
lower=trunc.lower[i,],upper=trunc.upper[i,]),simplify="array")
```

```
##  mult.obs x observations
moments1=t(sapply(1:observations, function(i) simulations[,i]$tmean,simplify="array"))
moments2=sapply(1:observations, function(i) simulations[,i]$tvar,simplify="array")

######### to calculate expectation of random effects
sigmab=sapply(1:observations, function(i) t(t(sigma.rand.old%*%t(z[,,i]))/
delta[i,])%*%solve(variance.ynew[,,i]),simplify="array")
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%((moments1[i,])-
mean.ynew[i,]), simplify="array")

if (dimension.sigma>1)
{ywave=sapply(1:observations, function(i) delta[i,]*(moments1[i,])-z[,,i]%*%
firstmomentb[,,i]+alpha[i,],simplify=F) }      else {ywave=sapply(1:observations,
function(i) delta[i,]*(moments1[i,])-z[,,i]%*%t(firstmomentb[i])+alpha[i,],simplify=F)}
#ywave is a list of vector column for each object in the study

betanew1=lm(do.call(cbind,ywave)[1,]~predictors.fixed1-1)$coef
#print(betanew1)
betanew2=lm(do.call(cbind,ywave)[2,]~predictors.fixed2-1)$coef
#print(betanew2)

#### update
predynew1=predictors.fixed1%*%betanew1
mean.ynew1=(predynew1-alpha1)/delta1
if(length(betaold2)>1) predynew2=predictors.fixed2%*%betanew2 else predynew2=
predictors.fixed2*betanew2
mean.ynew2=(predynew2-alpha2)/delta2
pred=cbind(predynew1,predynew2)
mean.ynew=cbind(mean.ynew1,mean.ynew2)
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%((moments1[i,])-
mean.ynew[i,]), simplify="array")

#######################################
################### delta estimates
##########################################
```

```
first=t(sapply(1:observations, function(i) diag(moments2[,,i])+(moments1[i,])^2,
simplify="array"))
## observations x mult.obs

if (dimension.sigma>1)
{pred.random=t(sapply(1:observations, function(i) z[,,i]%*%firstmomentb[,,i],
simplify=T)) }    else {pred.random=t(sapply(1:observations, function(i)
z[,,i]%*%t(firstmomentb[i]),simplify=T)) }

second1=moments1*pred

second2=sapply(1:observations, function(i)
sigmab[,,i]%*%(moments2[,,i]+moments1[i,]%*%t(moments1[i,])-mean.ynew[i,]%*%
t(moments1[i,])),simplify="array")
## for each i I have a q x n_i matrix - I need a column for each j
#### I need the columns in the last part to be y_ijnew*mean.ynew

if (dimension.sigma==1) second2=t(sapply(1:observations, function(i)
diag(z[,,i]%*%t(second2[,,i])), simplify="array")) else second2=
t(sapply(1:observations, function(i) diag(z[,,i]%*%second2[,,i]),simplify="array"))

third=moments1*delta-pred-pred.random

second3=-moments1*alpha

a1=sum(first[,1][data.ordinal1==2])+sum(n1[3])
a2=sum(first[,2][data.ordinal2==2])+sum(n2[3])

b1=sum((second1[,1]+second2[,1]+second3[,1])[data.ordinal1==2])-
sum(third[,1][data.ordinal1==3])
b2=sum((second1[,2]+second2[,2]+second3[,2])[data.ordinal2==2])-
sum(third[,2][data.ordinal2==3])

#c=-n[(k+1)]
c1=-n1[2]
```

```
c2=-n2[2]

deltanew1=(b1+sqrt(b1^2-4*a1*c1))/(2*a1)
#print(deltanew1)
deltanew2=(b2+sqrt(b2^2-4*a2*c2))/(2*a2)
#print(deltanew2)

## update

delta1=ifelse(data.ordinal1==1 |data.ordinal1==3,1,deltanew1)
delta2=ifelse(data.ordinal2==1 |data.ordinal2==3,1,deltanew2)
delta=cbind(delta1,delta2)
alpha1=ifelse(data.ordinal1==1 |data.ordinal1==2,0, deltanew1)
alpha2=ifelse(data.ordinal2==1 |data.ordinal2==2,0, deltanew2)
alpha=cbind(alpha1,alpha2)
mean.ynew1=(predynew1-alpha1)/delta1
mean.ynew2=(predynew2-alpha2)/delta2
pred=cbind(predynew1,predynew2)
mean.ynew=cbind(mean.ynew1,mean.ynew2)
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%((moments1[i,])-
mean.ynew[i,]), simplify="array")
sigmab=sapply(1:observations, function(i) t(t(sigma.rand.old%*%t(z[,,i]))/
delta[i,])%*%solve(variance.ynew[,,i]),simplify="array")

############################### sigma.rand estimate
if(dimension.sigma==1)
sigma.rand=sapply(1:observations, function(i)
sigma.rand.old-sigmab[,,i]%*%((z[,,i]%*%sigma.rand.old)/delta[i,])+
sigmab[,,i]%*%(moments2[,,i]+(moments1[i,])%*%t(moments1[i,])-
(moments1[i,])%*%t(mean.ynew[i,]) -(mean.ynew[i,])%*%t(moments1[i,])+
(mean.ynew[i,])%*%t(mean.ynew[i,]))%*%sigmab[,,i], simplify=T) else
sigma.rand=sapply(1:observations, function(i)
sigma.rand.old-sigmab[,,i]%*%((z[,,i]%*%sigma.rand.old)/delta[i,])+
sigmab[,,i]%*%(moments2[,,i]+(moments1[i,])%*%t(moments1[i,])-
(moments1[i,])%*%t(mean.ynew[i,])-(mean.ynew[i,])%*%t(moments1[i,])+
```

```
(mean.ynew[i,])%*%t(mean.ynew[i,]))%*%t(sigmab[,,i]), simplify=T)

if(dimension.sigma==1) sigma.rand.new=mean(sigma.rand) else sigma.rand.new=
matrix(apply(sigma.rand,1,mean),ncol=dimension.sigma)

#print(sigma.rand.new)
cat(number.it, " ")
#######################
#print(c(number.it,sigma.rand.new,betanew1,betanew2,deltanew1,deltanew2))
} # while

print(c(number.it,sigma.rand.new,betanew1,betanew2,deltanew1,deltanew2))
c(sigma.rand.new,betanew1,betanew2,deltanew1,deltanew2)

} #function

################### data simulation
library(MASS)
random.int=mvrnorm(n = 500, mu=c(0,0), Sigma=matrix(c(1,0.8,0.8,1),ncol=2))
l=length(random.int[,1])
int1=-0.5
int2=1
pred1=rnorm(l,1,0.5)
pred2=rexp(l,1)
slope1=1
slope2=-.5
y1=int1+slope1*pred1+random.int[,1]+rnorm(l,0,1)
y2=int2+slope2*pred2+random.int[,2]+rnorm(l,0,1)
par(mfrow=c(1,2))
hist(y1);hist(y2)

data.ordinal1=ifelse(y1<=0,1,ifelse(y1<=1.2,2,3))
data.ordinal2=ifelse(y2<=0,1,ifelse(y2<=.7,2,3))
table(data.ordinal1)
table(data.ordinal2)
```

```
predictors.random1=rep(1,l)
predictors.random2=predictors.random1
predictors.fixed1=cbind(1,pred1)
predictors.fixed2=cbind(1,pred2)

start.values.beta1=c(-0.5,1)
start.values.beta2=c(1,-0.5)
start.values.delta1=1.2
start.values.delta2=.7
start.values.sigma.rand=matrix(c(1,0.8,0.8,1),ncol=2)
#montecarlo=500

ecm.mult.ord(start.values.beta1,start.values.beta2,start.values.delta1,
start.values.delta2,start.values.sigma.rand,data.ordinal1,data.ordinal2,
predictors.fixed1,predictors.fixed2,predictors.random1,
predictors.random2,epsilon=0.0001)->proba0001
```

Приложение B: R код за ECM алгоритъм за корелиран “probit” модел за една наредена категорна и една нормална величини от дългосрочни проучвания

```
################################################################
condNormal <- function(x.given, mu, sigma, given.ind, req.ind){
# Returns conditional mean and variance of x[req.ind]
# Given x[given.ind] = x.given
# where X is multivariate Normal with
# mean = mu and covariance = sigma
#
B <- sigma[req.ind, req.ind]
C <- sigma[req.ind, given.ind, drop=FALSE]
D <- sigma[given.ind, given.ind]
CDinv <- C %*% solve(D)
cMu <- c(mu[req.ind] + CDinv %*% (x.given - mu[given.ind]))
cVar <- B - CDinv %*% t(C)
list(condMean=cMu, condVar=cVar)
}
```

```
##################################################
######################   ecm algorithm
#######################################################
ecm.ord.plus.cont.exact=function(start.values.beta.ordinal,
start.values.beta.continuous, start.values.delta,start.values.sigma.rand,
start.values.sigma22, start.values.lambda,data.ordinal,data.continuous,
predictors.fixed.ordinal,predictors.fixed.continous,
predictors.random.ordinal,predictors.random.continuous, epsilon) {

########################################
### sigma.rand is the covaraince matrix of the random effects
##################################

library(MASS)
library(tmvtnorm)

betanew.ordinal=start.values.beta.ordinal
betaold.ordinal=start.values.beta.ordinal-2
betanew.continuous=start.values.beta.continuous
betaold.continuous=start.values.beta.continuous-2

deltanew=start.values.delta
deltaold=start.values.delta-2

sigma.rand.new=start.values.sigma.rand
sigma.rand.old=start.values.sigma.rand+0.01

sigma22new=start.values.sigma22
sigma22old=start.values.sigma22+0.01
lambdanew=start.values.lambda
lambdaold=start.values.lambda+0.01

number.it=0

# definition of lower and upper boundary of the truncated normal distr.
```

```
trunc.lower=ifelse(data.ordinal==1,-Inf,0)
trunc.upper=ifelse(data.ordinal==1,0,ifelse(data.ordinal<3,1,Inf))
#number of observations in each category
n=0
for (i in 1:3) n[i]=sum(data.ordinal==i)

observations=length(data.ordinal[,1])
mult.obs=ncol(data.ordinal)
p=0
p[1]=ncol(predictors.random.ordinal)/mult.obs
#the dimension of the random effects for the ordinal response
p[2]=ncol(predictors.random.continuous)/mult.obs
#the dimension of the random effects for the continuous response
dimension.sigma=ifelse(length(sigma.rand.old)==1,1,ncol(sigma.rand.old))
#the dimension of the random effects
#######################################################################
############             while loop           #########################
##########################################################################

while(sum(abs(c(betanew.ordinal-betaold.ordinal,deltanew-deltaold,
betanew.continuous-betaold.continuous,sigma.rand.new-sigma.rand.old,
sigma22new-sigma22old,lambdanew-lambdaold))>epsilon)>0) {

number.it=number.it+1
#print(number.it)
betaold.ordinal=betanew.ordinal
deltaold=deltanew
betaold.continuous=betanew.continuous
sigma22old=sigma22new
lambdaold=lambdanew
sigma.rand.old=sigma.rand.new

#######################################################################
delta=ifelse(data.ordinal==1 |data.ordinal==3,1,deltaold)
alpha=ifelse(data.ordinal==1 |data.ordinal==2,0, deltaold)
```

```
##############################################
### marginal expectation of ynew #####
#####################################
predyneword=sapply(1:mult.obs, function(i) (predictors.fixed.ordinal[[i]]%*%
betaold.ordinal), simplify=TRUE)
## makes predyneword a matrix: observations x mult.obs
mean.yneword=(predyneword-alpha)/delta
mean.ycontinuous=sapply(1:mult.obs, function(i)
(predictors.fixed.continuous[[i]]%*%betaold.continuous), simplify=TRUE)
## makes mean.ycontinuous a mtarix: observations x mult.obs
mean.ynew=cbind(mean.yneword,mean.ycontinuous)
## observations x 2*mult.obs

z.ordinal=sapply(1:observations,function(i) matrix(predictors.random.ordinal[i,],
ncol=p[1], byrow=T), simplify="array")
## mult.obs x p[1] x observations
z.continuous=sapply(1:observations, function(i)
matrix(predictors.random.continuous[i,],ncol=p[2], byrow=T), simplify="array")
## mult.obs x p[2] x observations

varord=sapply(1:observations, function(i)
(z.ordinal[,,i]%*%matrix(sigma.rand.old[1:p[1],1:p[1]])%*%t(z.ordinal[,,i])+
diag(mult.obs)*(1+lambdaold^2*sigma22old))/(delta[i,]%*%t(delta[i,])),
simplify="array")
# mult.obs x mult.obs x observations

varordcont=sapply(1:observations, function(i)
(z.ordinal[,,i]%*%matrix(sigma.rand.old[1:p[1],(p[1]+1):dimension.sigma])%*%
t(z.continuous[,,i])+diag(mult.obs)*(lambdaold*sigma22old))/(delta[i,]),
simplify="array")

varcont=sapply(1:observations, function(i)
(z.continuous[,,i]%*%matrix(sigma.rand.old[(p[1]+1):dimension.sigma,
(p[1]+1):dimension.sigma])%*%t(z.continuous[,,i])+diag(mult.obs)*sigma22old),
```

```
simplify="array")
# mult.obs x mult.obs x observations

variance.ynew=sapply(1:observations, function(i)
rbind(cbind(varord[,,i],varordcont[,,i]),cbind(t(varordcont[,,i]),varcont[,,i])),
simplify="array")
##2*mult.obs x 2*mult.obs x observations


#################### conditional distribution of y latent given observed normal
latent=sapply(1:observations, function(i)
condNormal(x=data.continuous[i,], mu=mean.ynew[i,], sigma=variance.ynew[,,i],
req=1:mult.obs, given=(mult.obs+1):(2*mult.obs)), simplify="array")

### computation of exact moments
simulations.exact=sapply(1:observations,function(i) mtmvnorm(mean=latent[,i]$condMean,
sigma=matrix(as.vector(latent[,i]$condVar),ncol=mult.obs),
lower=trunc.lower[i,],upper=trunc.upper[i,]),simplify="array")
##  mult.obs x observations
moments1=t(sapply(1:observations, function(i)
simulations.exact[,i]$tmean,simplify="array"))
moments2=sapply(1:observations, function(i)
simulations.exact[,i]$tvar,simplify="array")
## the variance of y latent

########## to calculate the expectation of random effects
covyordinalb=sapply(1:observations, function(i)
matrix(z.ordinal[,,i],ncol=p[1])%*%sigma.rand.old[1:p[1],]/delta[i,],simplify="array")
## mult.obs x dimension.sigma x observations

covycontb=sapply(1:observations, function(i)
matrix(z.continuous[,,i], ncol=p[2])%*%
sigma.rand.old[(p[1]+1):dimension.sigma,],simplify="array")
#mult.obs x dim.sigma x observations
```

```
covynewb=sapply(1:observations, function(i)
rbind(covyordinalb[,,i],covycontb[,,i]), simplify="array")
#2*mult.obs x dimension.sigma x observations

sigmab=sapply(1:observations, function(i)
t(covynewb[,,i])%*%solve(variance.ynew[,,i]), simplify="array")
#dimension.sigma x 2*mult.obs x observations

firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%
(c(moments1[i,],data.continuous[i,])-mean.ynew[i,]), simplify="array")
#### dimension.sigma x 1 x observations
##############################################
mu2=t(sapply(1:observations, function(i) mean.ycontinuous[i,]+
z.continuous[,,i]*firstmomentb[,,i][(p[1]+1):sum(p)], simplify="array"))
# observations x mult.obs
#######################################################################
#########################  beta.ordinal estimates
#######################################################################
ywave=sapply(1:observations, function(i) delta[i,]*moments1[i,]-z.ordinal[,,i]*
firstmomentb[,,i][1:p[1]]+alpha[i,]-lambdaold*(data.continuous[i,]-mu2[i,]),
simplify=F)
#ywave is a list of vector column for each object in the study
betanew.ordinal=lm(c(t(do.call(cbind,ywave)))~
do.call(rbind,predictors.fixed.ordinal)-1)$coef
#print(betanew.ordinal)

#### update
predyneword=sapply(1:mult.obs, function(i) (predictors.fixed.ordinal[[i]]%*%
betanew.ordinal), simplify=TRUE)
mean.yneword=(predyneword-alpha)/delta
mean.ynew=cbind(mean.yneword,mean.ycontinuous)
firstmomentb=sapply(1:observations, function(i) sigmab[,,i]%*%
(c(moments1[i,],data.continuous[i,])-mean.ynew[i,]), simplify="array")

#################################################################
```

```
##############################        sigma22 estimate
#######################################################################

# the variance of ynew given observed data
zero=sapply(1:observations, function(i)  {
a=matrix(0,ncol=2*mult.obs,nrow=2*mult.obs); a[1:mult.obs,1:mult.obs]=
moments2[,,i]; a},simplify="array")
## 2*mult.obs x 2*mult.obs x observations

## the second moment of random effects given observed data: E(b_i*b_i')
sigma.rand=sapply(1:observations, function(i)
sigma.rand.old-sigmab[,,i]%*%covynewb[,,i]+
sigmab[,,i]%*%(zero[,,i]+c(moments1[i,],data.continuous[i,])%*%
t(c(moments1[i,],data.continuous[i,]))-c(moments1[i,],data.continuous[i,])%*%
t(mean.ynew[i,])-mean.ynew[i,]%*%t(c(moments1[i,],data.continuous[i,]))+
(mean.ynew[i,])%*%t(mean.ynew[i,]))%*%t(sigmab[,,i]), simplify=T)
#### dimension.sigma^2 x observations

### expectation of mu2 squared
mu22=t(sapply(1:observations, function(i) mean.ycontinuous[i,]*
(mean.ycontinuous[i,]+2*z.continuous[,,i]*firstmomentb[,,i][(p[1]+1):sum(p)])+
z.continuous[,,i]^2*sigma.rand[4,i], simplify="array"))

sigma22new=sum(data.continuous*(data.continuous-2*mu2)+mu22)/(mult.obs*observations)
#print(sigma22new)


#######################################################################
######################        lambda estimate
#######################################################################
randompart.ordinal=t(sapply(1:observations, function(i) z.ordinal[,,i]*
firstmomentb[1:p[1],,i], simplify=TRUE))
randompart.continuous=t(sapply(1:observations, function(i) z.continuous[,,i]*
firstmomentb[(p[1]+1):dimension.sigma,,i], simplify=TRUE))
```

```
meanylatent=(predyneword+randompart.ordinal-alpha)/delta
## mean of conditional expectation of y latent on random effects given observed data

z=sapply(1:observations, function(i) rbind(cbind(z.ordinal[,,i],
matrix(0,ncol=p[2],nrow=mult.obs)),cbind(matrix(0,ncol=p[1],
nrow=mult.obs),z.continuous[,,i])), simplify="array")
### mult.obs x dimension.sigma x observations

gadost1=sapply(1:observations, function(i)
moments2[,,i]+moments1[i,]%*%t(moments1[i,])-mean.yneword[i,]%*%
t(moments1[i,]), simplify="array")
##mult.obs x mult.obs x observations
gadost2=sapply(1:observations, function(i)
(data.continuous[i,]-mean.ycontinuous[i,])%*%t(moments1[i,]),simplify="array")
##mult.obs x mult.obs x observations
gadost=sapply(1:observations, function(i)
rbind(gadost1[,,i],gadost2[,,i]), simplify="array")
## 2mult.obs x mult.obs x observations

## the expectation of z_i*b_i*y_1i_new'
gadostbig=sapply(1:observations, function(i)
z[,,i]%*%sigmab[,,i]%*%gadost[,,i], simplify="array")
## 2mult.obs x mult.obs x observations

gadostbigb1=t(sapply(1:observations, function(i)
diag(gadostbig[,,i]), simplify="array"))

gadostbigb2=t(sapply(1:observations, function(i)
diag(gadostbig[,,i][(mult.obs+1):(2*mult.obs),]), simplify="array"))

sigma.rand=sapply(1:observations, function(i) matrix(sigma.rand[,i],
ncol=sum(p)), simplify="array")

expectationb1ib2i=t(sapply(1:observations, function(i)
diag(z.ordinal[,,i]%*%matrix(sigma.rand[1:p[1],(p[1]+1):sum(p),i],
```

```
ncol=sum(p)-p[1])%*%t(z.continuous[,,i])), simplify="array"))
##   mult.obs x mult.obs
## the first column is z[,,i]*e(b1ib2i)*z[1,,i]
## i need only the diagonal element !!!!!!! ###

lambdanew=sum( delta*(moments1-meanylatent)*(data.continuous-mean.ycontinuous)-
randompart.continuous*(alpha-predyneword)+expectationb1ib2i-delta*
gadostbigb2)/sum(data.continuous*(data.continuous-2*mu2)+mu22)
#print(lambdanew)

#################################################################
###################                delta estimates
#################################################################
first=t(sapply(1:observations, function(i) diag(moments2[,,i])+
(moments1[i,])^2, simplify="array"))
## observations x mult.obs

second=moments1*(predyneword-alpha+lambdanew*(data.continuous-
mean.ycontinuous))+gadostbigb1-lambdanew*gadostbigb2

third=moments1*delta-predyneword-randompart.ordinal-lambdanew*
(data.continuous-mean.ycontinuous-randompart.continuous)

a=sum(first[data.ordinal==2])+sum(n[3])

b=sum(second[data.ordinal==2])-sum(third[data.ordinal==3])

c=-n[2]

deltanew=(b+sqrt(b^2-4*a*c))/(2*a)
#print(deltanew)

#### update
delta=ifelse(data.ordinal==1 |data.ordinal==3,1,deltanew)
alpha=ifelse(data.ordinal==1 |data.ordinal==2,0, deltanew)
```

```
mean.yneword=(predyneword-alpha)/delta
mean.ynew=cbind(mean.yneword,mean.ycontinuous)

########## to calculate the expectation of random effects
covyordinalb=sapply(1:observations, function(i)
matrix(z.ordinal[,,i],ncol=p[1])%*%sigma.rand.old[1:p[1],]/delta[i,],simplify="array")
## mult.obs x dimension.sigma x observations

covynewb=sapply(1:observations, function(i)
rbind(covyordinalb[,,i],covycontb[,,i]), simplify="array")
#2*mult.obs x dimension.sigma x observations

sigmab=sapply(1:observations, function(i)
t(covynewb[,,i])%*%solve(variance.ynew[,,i]), simplify="array")
#dimension.sigma x 2*mult.obs x observations

firstmomentb=sapply(1:observations, function(i)
sigmab[,,i]%*%(c(moments1[i,],data.continuous[i,])-mean.ynew[i,]), simplify="array")

##############################################################################
######################                 beta continuous
#############################################################################
term=1+lambdanew^2*sigma22new
sumsum=term*(data.continuous-randompart.continuous)-lambdanew*
sigma22new*(delta*moments1+alpha-predyneword-randompart.ordinal)
vect=apply(sapply(1:mult.obs, function(j) apply(sumsum[,j]*
predictors.fixed.continuous[[j]],2,sum),simplify=T),1,sum)

m=apply(sapply(1:mult.obs, function(j) apply(apply(predictors.fixed.continuous[[j]],
1, function(i) i%*%t(i)),1,sum), simplify="array"),1,sum)
m=matrix(m, ncol=ncol(predictors.fixed.continuous[[1]]),byrow=F)
m=term*m

betanew.continuous=solve(m,vect)
#print(betanew.continuous)
```

```
#### update
mean.ycontinuous=sapply(1:mult.obs, function(i) (predictors.fixed.continuous[[i]]%*%
betanew.continuous), simplify=TRUE)
mean.ynew=cbind(mean.yneword,mean.ycontinuous)

#######################################################################
########################           sigma.rand estimate
#######################################################################
sigma.rand=sapply(1:observations, function(i)
sigma.rand.old-sigmab[,,i]%*%covynewb[,,i]+
sigmab[,,i]%*%(zero[,,i]+c(moments1[i,],data.continuous[i,])%*%t(c(moments1[i,],
data.continuous[i,]))-c(moments1[i,],data.continuous[i,])%*%t(mean.ynew[i,])-
mean.ynew[i,]%*%t(c(moments1[i,],data.continuous[i,]))+(mean.ynew[i,])%*%
t(mean.ynew[i,]))%*%t(sigmab[,,i]), simplify=T)

#if(dimension.sigma==1) sigma.rand.new=mean(sigma.rand) else
sigma.rand.new=matrix(apply(sigma.rand,1,mean),ncol=dimension.sigma)

#print(sigma.rand.new)

########################
cat("number.it: ",number.it,"sigma.rand.new: ",sigma.rand.new,"betanew.ordinal: ",
betanew.ordinal,"\n","betanew.continuous: ",betanew.continuous,"deltanew: ",
deltanew,"lambdanew: ",lambdanew,"sigma22new: ",sigma22new,"\n")
print(Sys.time())
} # while

cat("Final estimates with accuracy=",epsilon," :","\n","number.it: ",number.it,
"sigma.rand.estimate: ",sigma.rand.new,"betanew.ordinal.estimate: ",
betanew.ordinal,"\n","betanew.continuous.estimate: ",betanew.continuous,
"deltanew.estimate: ",deltanew,"lambdanew.estimate: ",lambdanew,
"sigma22new.estimate: ",sigma22new,"\n")
```

```
c(sigma.rand.new,betanew.ordinal,betanew.continuous,deltanew,lambdanew,sigma22new)

} #function

################### data simulation
library(MASS)
random.int=mvrnorm(n = 2500, mu=c(0,0), Sigma=matrix(c(1,-0.8,-0.8,1),ncol=2))
l=length(random.int[,1])
mult.obs=3
random.error=sapply(1:mult.obs, function(i) mvrnorm(n = l, mu=c(0,0),
Sigma=matrix(c(4/3,2/sqrt(3),2/sqrt(3),4),ncol=2)),
simplify="array")
#random.int=rnorm(500,0,0.1)
int1=-0.5
int2=1
slope1=1
slope2=-.5
time=matrix(rep(1:mult.obs,l),byrow=T,ncol=mult.obs)
y1=int1+slope1*time+random.int[,1]+random.error[,1,]
y2=int2+slope2*time+random.int[,2]+random.error[,2,]

data.ordinal=ifelse(y1<=0,1,ifelse(y1<=1.2,2,3))
data.continuous=y2
table(data.ordinal)

predictors.random.ordinal=matrix(rep(1,l*mult.obs),ncol=mult.obs)
predictors.random.continuous=predictors.random.ordinal
predictors.fixed.ordinal=sapply(1:mult.obs, function(i) cbind(1,time[,i]), simplify=F)
predictors.fixed.continuous=predictors.fixed.ordinal

start.values.beta.ordinal=c(-0.5,1)
start.values.beta.continuous=c(1,-0.5)
start.values.delta=1.2

start.values.sigma.rand=matrix(c(1,-0.8,-0.8,1),ncol=2)
```

```
start.values.sigma22=4
start.values.lambda=1/(2*sqrt(3))

ecm.ord.plus.cont.exact(start.values.beta.ordinal,start.values.beta.continuous,
start.values.delta,start.values.sigma.rand,start.values.sigma22, start.values.lambda,
data.ordinal,data.continuous,predictors.fixed.ordinal,predictors.fixed.continous,
predictors.random.ordinal,predictors.random.continuous, epsilon=0.001)->proba0002
```

Библиография

Б. Димитров и Н. Янев. *Теория на вероятностите и математическа статистика.* Университетско издателство "Св. Климент Охридски", София, Второ издание, 1998. 11

J. Aitchison and S. D. Silvey. The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44(1-2):131–140, 1957. doi: 10.1093/biomet/44.1-2.131. URL http://biomet.oxfordjournals.org/content/44/1-2/131.short. 3

C. N. Andreassen. Can risk of radiotherapy-induced normal tissue complications be predicted from genetic profiles? *Acta Oncol*, 44:801–815, 2005. 44

C. N. Andreassen, J. Alsner, and J. Overgaard. Prediction of normal tissue radiosensitivity from polymorphisms in candidate genes. *Radiother Oncol*, 69:127–135, 2003. 44

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. 3

C. I. Bliss. The method of probits. *Science*, pages 38–39, 1934a. 3

C. I. Bliss. The method of probits - a correction. *Science*, pages 409–410, 1934b. 3

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993. 5, 15

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. 23, 27

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. 4

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. 18, 30

K. De Ruyck, M. Van Eijkeren, K. Claes, R. Morthier, A. De Paepe, A. Vral, L. De Ridder, and H. Thierens. Radiation-induced damage to normal tissues after radiotherapy in patients treated for gynecologic tumors: association with single nucleotide polymorphisms in XRCC1, XRCC3, and OGG1 genes and in vitro chromosomal radiosensitivity in lymphocytes. *Int J Radiat Oncol Biol Phys*, 62: 1140–1149, 2005. 44

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. 5, 14, 18

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. 65

P. J. Diggle and M. G. Kenward. Informative dropout in longitudinal data analyses (with discussion). *Applied Statistics*, 43:49–93, 1994. 65

P. J. Diggle, K.-L. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, New York, 1996. 2, 10, 13

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. 30

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. 5

A. Edwards and L. Thurstone. An internal concistency check for scale values determined by the method of successive intervals. *Psychometrica*, 17(2):169–180, 1952. 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. 27, 39, 58

R. M. Elashoff, G. Li, and N. Li. An Approach to Joint Analysis of Longitudinal Measurements and Competing Risks Failure Time Data. *Statistics in Medicine*, 26: 2813–2835, 2007. 66

R. M. Elashoff, G. Li, and N. Li. A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, 64:762–771, 2008. 66

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, second edition, 2001. 3, 4, 5

Julian J. Faraway. *Linear Models with R*. Chapman & Hall/CRC, 2004. URL http://www.maths.bath.ac.uk/~jjf23/LMR/. ISBN 1-584-88425-8. 9

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. 2

J. H. Gaddum. Methods of biological assay depending on a quantal response. *Reports on biological standards. III.*, 1933. 3

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. 5

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. 4

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. 40, 47

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. 16

D. Grigorova and R. Gueorguieva. EM Algorithms for MLE of Correlated Probit Models. In *[MIE 2013] Proceedings*, pages 17–24, 2013b. 31

L. Grilli and C. Rampichini. Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics*, 28:31–44, 2003. 4, 5

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. 66

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. 4

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. 5

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. 4, 5, 62

X. Guo and B. P. Carlin. Separate and Joint Modelling of Longitudinal and Event Time Data Using Standard Computer Packages. *The American Statistician*, 58:16–24, 2004. 65

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480, 2000. 65

J. W. Hogan, J. Roy, and C. Korkontzelou. Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Statistics in Medicine*, 23:1455–1497, 2004. 65

H. Kawakatsu and A. G. Largey. EM algorithms for ordered probit models with endogenous regressors. *Econometrics Journal*, 12:164–186, 2009. 4, 6, 18, 33, 50

J. P. Klein. Semiparametric estimation of random effects using the cox model based on em algorithm. *Biometrics*, 48:795–806, 1992. 66

M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 2005. 3, 9

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. 2

Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–678, 1996. 8

K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13. 2

R. J. A. Little and Y. Wang. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, 52:98–111, 1996. 65

C. Liu, D. Rubin, and Y. Wu. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85(4):755–770, 1998. 30, 47, 63

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. 4, 14

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. 30, 47, 63

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. 23, 38

P. McCullagh and J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989. 4, 12

C. McCulloch. Joint modelling of mixed outcome types using latent variables. *Stat Methods Med Res*, 17(1):53–73, 2008. 5

C. A. McGilchrist. REML estimation for survival models with frailty. *Biometrics*, 49: 221–225, 1993. 66

C. A. McGilchrist and C. W. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47:461–466, 1991. 66

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. 25, 37, 56

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. 5, 24

G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. A. Sorensen. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian J. Stat.*, 19:25–43, 1992. 66

Y. Ochi and R. L. Prentice. Likelihood inference in a correlated probit regression model. *Biometrika*, 71(3):531–543, 1984. doi: 10.1093/biomet/71.3.531. URL http://biomet.oxfordjournals.org/content/71/3/531.abstract. 3

H. D. Patterson and Thompson. Recovery of interblock information when block sizes are unequal. *Biometrika*, 58:545–554, 1971. 11

J. C. Pinheiro and E. C. Chao. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):58–81, 2006. doi: 10.1198/106186006X96962. URL http://www.tandfonline.com/doi/abs/10.1198/106186006X96962. 14

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. iii, 6, 26, 30, 47, 63

S. Ripatti and J. Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56:1016–1022, 2000. 66

P. A. Ruud. Extensions of estimation methods using the EM algorithm. *Journal of Econometrics*, 49(3):305–341, September 1991. 18, 33, 50

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. 66

T. M. Therneau and P. M. Grambsch. *Modeling survival data Extending the Cox model*. Springer-Verlag, New York, 2001. 65

A. A. Tsiatis and M. Davidian. Joint Modelling of Longitudinal and Time to Event Data: An Overview. *Statistica Sinica*, 14:809–834, 2004. 65

P. J. M. Verweij and H. C. Van Houwelingen. Penalized likelihood in cox regression. *Statistics in Medicine*, 13:2427–2436, 1994. 66

Robert E. Weiss. *Modeling Longitudinal Data*. Springer Science+Business Media, New York, 2005. 2, 10, 29

S. Wilhelm. Gibbs sampler for the truncated multivariate normal distribution. Electronic, April 6 2012. http://cran.r-project.org/web/packages/tmvtnorm/vignettes/GibbsSampler.pdf. 27

S. Wilhelm and B. G. Manjunath. *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*, 2012. URL `http://CRAN.R-project.org/package=tmvtnorm`. R package version 1.4-7. 27, 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. 5

M. C. Wu and R. J. Caroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44: 175–188, 1987. 65

Декларация за оригиналност на резултатите

Декларирам, че настоящата дисертация на тема „ЕМ 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.

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

16.12.2013 г., Подпис:

София Деница Григорова