

UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Trendy ve výskytu Downova syndromu



Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **Mgr. Ondřej Vencálek, Ph.D.**
Vypracoval(a): **Bc. Pavla Stará**
Studijní program: B1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Pavla Stará

Název práce: Trendy ve výskytu Downova syndromu

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Cílem práce je ověřit či vyvrátit hypotézu růstu rizika výskytu Downova syndromu v české populaci v letech 1961-2010. Ve zmíněném období došlo ke zvyšování průměrného věku matek. Věk matky je přitom pro určení rizika důležitým faktorem (s rostoucím věkem roste riziko). Zároveň však došlo k rozvoji a rozšíření prenatální diagnostiky. Tyto změny měly významný vliv na počet narozených jedinců s Downovým syndromem.

Klíčová slova: Downův syndrom, prenatální diagnostika, riziko, logistická regrese, joinpoint regrese, standardizace

Počet stran: 82

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Pavla Stará

Title: Trends in the incidence of Down's syndrome

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2016

Abstract: The aim is to verify or disprove the hypothesis increase risk of Down's syndrome in population of the Czech Republic in the years 1961-2010. In the mentioned period, it recorded an increase in maternal age. However maternal age are an important factor in determining risk (with increasing age increases risk). At the same time, there was also the development and expansion of prenatal diagnosis. These changes have had a significant impact on the number of individuals born with Down's syndrome.

Key words: Down's syndrome, prenatal diagnosis, risk, logistic regression, joinpoint regression, standardization

Number of pages: 82

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Mgr. Ondřeje Vencálka, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne

.....

podpis

Obsah

Úvod	11
1 Motivace	13
2 Downův syndrom	15
2.1 Historie	15
2.2 Charakteristika Downova syndromu	16
2.3 Prenatální diagnostika	17
2.3.1 Neinvazivní metody prenatální diagnostiky	18
2.3.2 Invazivní metody prenatální diagnostiky	19
2.3.3 Incidence a prevalence	21
3 Seznámení se s daty	23
3.1 Vstupní data	23
3.2 Grafy	24
4 Riziko výskytu Downova syndromu	35
4.1 Technický problém	35
4.2 Výpočet rizika výskytu Downova syndromu	38
5 Model závislosti rizika Downova syndromu na věku matky	41
5.1 Logistická regrese	43
5.1.1 Metoda maximální věrohodnosti	45
5.2 Joinpoint regrese	46
5.3 Model	49
5.3.1 Řešení pomocí logistické regrese	49
5.3.2 Řešení pomocí joinpoint regrese	56
6 Standardizace	61
6.1 Obecný popis procesu standardizace	61
6.1.1 Přímá standardizace	62
6.2 Klouzavý průměr	64
6.3 Aplikace na námi analyzovaná data	68

Závěr	79
Literatura	81

Seznam obrázků

2.1	Trisomie chromosomu 21 [21]	17
3.1	Struktura dat	23
3.2	Rozvoj prenatalní diagnostiky	25
3.3	Celkový počet narozených Downů	26
3.4	Počet narozených dětí v ČR (1961-2010)	27
3.5	Průměrný věk matek (1961-2010)	28
3.6	Věkové skupiny matek	28
3.7	Incidence DS (1961-2010)	33
3.8	Efektivita prenatalní diagnostiky	34
4.1	Podíl Downů s neznámým věkem matky	36
4.2	Riziko výskytu Downova syndromu	38
4.3	Logit rizika výskytu Downova syndromu	40
5.1	Modely závislosti rizika DS na věku matky I. [11]	42
5.2	Modely závislosti rizika DS na věku matky II. [11]	43
5.3	Model, prostřednictvím něhož je iterativně odhadnuta joinpoint regrese pomocí funkce segmented. [19]	48
5.4	Model závislosti rizika DS na věku matky	52
5.5	Tabulka - předpokládaná prevalence narozených Downů na 1 000 novorozenců (zahraniční referenční studie) [11]	54
5.6	Porovnání rizika podle věku matky našeho modelu a referenčních zahraničních studií (maximum a minimum)	55
5.7	Porovnání logitu rizika podle věku matky našeho modelu a referenčních zahraničních studií (maximum a minimum)	55
5.8	Výsledek funkce slope a summary pro joinpoint model s jedním bodem zvratu	59
5.9	Odhadnutý model vypočítaný v programu R. Znázorňuje logit pravděpodobnosti výskytu Downova syndromu v závislosti na věku matky.	60
6.1	Vývoj standardizovaných počtů narozených DS (1961-2010)	73
6.2	Vývoj standardizovaných počtů diagnostikovaných (prenatálně i postnatálně) DS (1961-2010)	74

6.3	Vývoj standardizovaných počtů narozených DS (1961-2010) - vyhlazení klouzavým průměrem	77
6.4	Vývoj standardizovaných počtů diagnostikovaných (prenatálně i postnatálně) DS (1961-2010) - vyhlazení klouzavým průměrem	78

Seznam tabulek

3.1 Incidence Downova syndromu v ČR	32
--	-----------

Poděkování

Ráda bych poděkovala svému vedoucímu diplomové práce, panu Mgr. Ondřejovi Vencálkovi, Ph.D., za jeho ochotu, vstřícný přístup, odbornou pomoc, poskytnutí cenných rad a čas, který mi věnoval při zpracování mé diplomové práce. Také bych chtěla vyjádřit svůj velký dík mé rodině a mému příteli, kteří mě během studia plně podpořovali a především motivovali.

Úvod

Tato diplomová práce byla vypracována jako završení dvouletého navazujícího studia magisterského studijního oboru Aplikace matematiky v ekonomii. Práce pojednává o Downově syndromu, což je vrozená chromosomální vada a postižený jedinec si s sebou celý život nese specifická postižení. Diplomová práce klade za cíl vysvětlit, zda nárůst počtu jedinců postižených Downovým syndromem je zapříčiněn zvyšujícím se věkem matek nebo zda dochází k dlouhodobým trendům v incidenci této vrozené vady.

Sledovali jsme vývoj za období 1961 – 2010 v populaci českých matek. V tomto období došlo ke zvyšování průměrného věku matek. Věk matky je zásadní faktor, jelikož s rostoucím věkem matky roste riziko výskytu Downova syndromu. Zároveň však v tomto období došlo k rozvoji a rozšíření prenatalní diagnostiky, která dokáže zachytit více těhotenství postižených Downovým syndromem. Tyto změny měly zásadní vliv na tom, kolik se narodilo jedinců s Downovým syndromem.

Práce je rozdělena do šesti kapitol. První, úvodní, kapitola je jakousi motivací a krátkým zamyšlením se nad dnešním světem, ve kterém žijeme. Ve druhé kapitole jsou shrnuty poznatky o Downovu syndromu. Zmíníme se o jeho historii, popíšeme si typické projevy tohoto postižení a seznámíme se se současnými metodami prenatalní diagnostiky. Třetí kapitola obsahuje popis obdržných dat a zároveň budeme získané informace, které nám data přináší, vykreslovat do grafů. Ve čtvrté kapitole krátce pojednáme o riziku výskytu Downova syndromu. Zde narážíme na první technický problém, který bude třeba vyřešit. Obsahem páté kapitoly je sestavení modelu závislosti rizika na věku matky. Nejdříve si popíšeme metodu logistické regrese a metodu joinpoint regrese, při které se využívá v programu R knihovna `segmented`.

Pak tyto metody aplikujeme na námi analyzovaná data. V poslední kapitole, kapitole páté, využijeme standardizaci, abychom se mohli oprostít od závislosti na věku matky a zjistit, jestli riziko výskytu Downova syndromu roste i přes tuto skutečnost. Pro lepší čitelnost grafů použijeme vyhlazení klouzavým průměrem. Ke zpracování výpočtů a obrázků bude použit program R a Excel.

Tato práce je psána za předpokladu, že čtenář má základní povědomí ze všech použitých matematických disciplín, především pak z pravděpodobnosti a statistiky a problematiky časových řad.

Práce je sázena v programu \LaTeX .

Kapitola 1

Motivace

Než se začneme zabývat samotným tématem diplomové práce, krátce se zamysleme, v jakém světě momentálně žijeme.

V první řadě žijeme v době, která je velice uspěchaná, plná stresu a honu za penězi. Lidé se štvou za budováním kariéry, chtějí vlastnit luxusní auta a přepychové domy, touží jezdit na drahé dovolené a užívat si života. S tím úzce souvisí plánování rodiny. Je docela běžné, že manželé či partneři zakládají rodinu až v pozdějším věku. Tudiž věk matek se zvyšuje. Mnozí čtenáři této práce si určitě říkají, že na tom není nic neobvyklého, že dnes se prostě „rodí“ později. To ano, ale napadlo nás někdy, že i přes tento fakt je naše tělo pořád zhruba stejné, jako tomu bylo před deseti, dvaceti lety? A že tento trend se může podepisovat na tom, že přibývá jak diagnostikovaných tak narozených dětí, které jsou nějak postižené (třeba Downovým syndromem)?

Dále se dnešní svět vyznačuje lidskou nevráživostí, na mnoha místech zeměkoule panují neustálé konflikty, rozbroje či války. Mnozí lidé trpí depresemi, jsou neustále nespokojení a touží po novém „lepší“ životě. Naše planeta se nám dočista mění před očima. Není to však náhodou dílem samotného člověka, který bezohledně ničí naši planetu? Příroda je mocná a umí se bránit tím, že nám špatné konání lidí začíná vracet. Velkým problémem se stává i zvyšující se znečištění a jeho dopad na zdraví člověka.

Z druhého úhlu pohledu je naopak dnešní doba poměrně vyspělá. Pokrok jde neustále kupředu, technologie jsou čím dál vyspělejší. Co se nám dříve zdálo nepředstavitelné, je dnes zcela běžné. Důkazem je toho prenatalní diagnostika, která je momentálně na velmi vysoké úrovni.

Downův syndrom je součástí Národního registru vrozených vad České republiky (ČR) a na území ČR je sledován dlouhodobě. Tento registr má u nás dlouholetou tradici a v současné době představuje nepřetržitou řadu od roku 1961. Již od tohoto roku jsou k dispozici výsledky registrace vrozených vad z celé naší republiky. Jako zajímavost uved' me, že Česká republika se stala jedním z prvních států na světě, kde byla zavedena registrace vrozených vad. [3]

Vývoj od roku 1961-2010 prošel zásadními změnami. Ve zmiňovaném období došlo k rozvoji a rozšíření prenatalní diagnostiky, která dokáže zachytit více případů těhotenství, která jsou postižená Downovým syndromem. V dnešní době je drtivá většina jedinců postižená Downovým syndromem (Downů) diagnostikována prenatalně. Druhou důležitou skutečností je, že došlo ke zvyšování průměrného věku matek a jak je všeobecně známo, věk matky je pro určení rizika velmi důležitým faktorem, neboť s rostoucím věkem matky roste riziko výskytu Downova syndromu. Tudíž můžeme vidět, že vzrůstá počet prenatalně diagnostikovaných plodů s Downovým syndromem. Všechny tyto změny mají významný vliv na tom, kolik se narodí Downů.

V této diplomové práci se budeme snažit vysvětlit, zda nárůst počtu případů Downova syndromu je zapříčiněn pouze vyšším věkem matek nebo zda dochází k dlouhodobým trendům v incidenci této vrozené vady.

Kapitola 2

Downův syndrom

V této kapitole se krátce seznámíme s tím, co je Downův syndrom (DS). Představíme si metody prenatalní diagnostiky, které jsou dnes používány v České republice, a ukážeme si, jaký význam představuje prenatalní diagnostika pro současnou populaci. Při tvorbě této kapitoly bylo čerpáno především z [3], [4], [16], [18], [21], [23], [20].

2.1. Historie

Je velice pravděpodobné, že s charakteristickými rysy Downova syndromu se setkávala řada lékařů již v dávné historii lidstva. Už od pradávna se mezi lidmi vyskytovaly bytosti, které se vymykaly běžným kritériím. Otázka, která nás určitě napadá, je, zda se Downův syndrom vyskytoval už na počátku civilizace nebo zda se jedná o jev dnešní moderní doby. Na tuto otázku nemůžeme najít přesnou odpověď, ale je vysoce pravděpodobné, že mnohé geneticky podmíněné nemoci a různé chromosomální abnormality se vyskytovaly už v předchozích tisíciletích. K přesnějšímu popisu došlo až s rozvojem lékařství.

Jako první v roce 1838 popsal lékař Jean Esquirol vzhled dítěte, jehož znaky odpovídají dnešnímu fenotypu¹ Downova syndromu. Ale až o několik let později, roku 1866 byl Downův syndrom uznán jako samostatná diagnóza a pojmenován po svém objeviteli, anglickém lékaři, Johnu Langdonu Downovi. Ten poprvé popsal charakteristické rysy syndromu, avšak zatím nebyly známy příčiny diagnózy. V roce 1932

¹Fenotyp představuje soubor všech pozorovatelných vlastností a znaků živého organismu.

lékař Waardenburg přišel s myšlenou, že příčinou Downova syndromu by mohla být chromosomální abnormalita. Trvalo však nějakou dobu, než jeho návrh byl potvrzen. V roce 1959, francouzský genetik a pediatr, Jérôme Lejeune nazval tuto abnormalitu jako trisomie chromosomu 21.

2.2. Charakteristika Downova syndromu

Downův syndrom je vrozená chromosomální vada, kdy postižený jedinec má v každé buňce svého těla jeden chromosom² navíc, to se označuje jako trisomie 21. Zdravý jedinec má 46 chromosomů uspořádaných ve 23 párech, nemocní mají 47 chromosomů – 22 párů a jednu trojici (obrázek 2.1). K této abnormalitě dochází těsně po oplodnění, kdy pohlavní buňky se podrobují zvláštnímu dělení nazývané jako nondisjunkce (chybného rozestupu chromosomů v průběhu I. či II. meiotického dělení)³. Může se vyskytnout i tzv. translokační forma trisomie 21, která je však velmi vzácná a u které je nadbytečný chromosom 21 připojen na jiný chromosom.

Četnost výskytu Downova syndromu je úzce spojená s věkem matky, protože žena má vajíčka ve vaječnicích uložena již od narození a vajíčka stárnou spolu se ženou. A proto čím jsou starší, tím jsou náchylnější na vznik genetické vady. Výskyt Downova syndromu je nahodilý a jeho přesné příčiny se do dnešní doby nepodařilo uspokojivě vysvětlit. Jako prevence proti narození postiženého dítěte je v době těhotenství prenatální diagnostika.

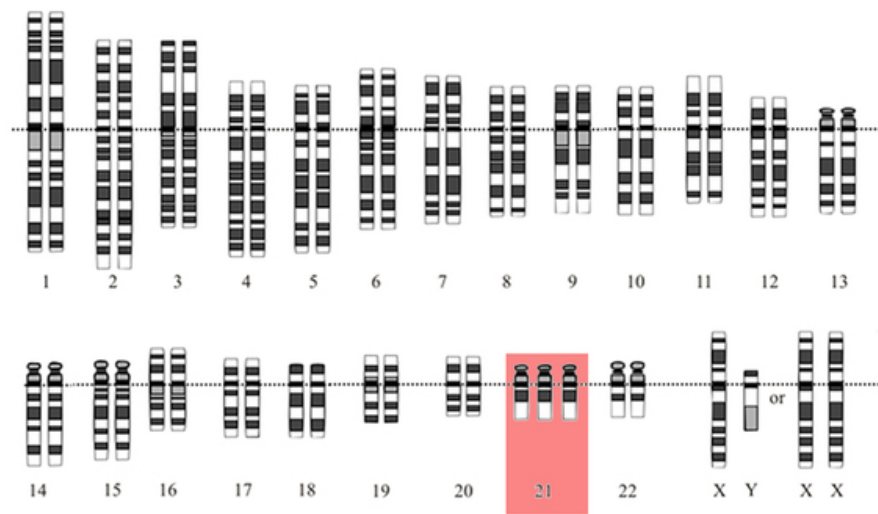
Downův syndrom je velmi dobře rozpoznatelný lidským okem. Projevuje se především mentální retardací středního až těžkého stupně. Lidé s Downovým syndromem trpí tělesnými malformacemi a mají mongoloidní vzhled. Tedy jsou například menší postavy, mají ploché rysy obličeje, atypický tvar hrudníku a mnoho dalších odlišností. K typickým vnitřním vadám například patří hlavně srdeční vady, mají sníženou imunitu a mají vyšší pravděpodobnost výskytu některých nemocí než běžná populace –

²Chromosomy jsou stočená vlákna deoxyribonukleové kyseliny (DNA), které nesou genetickou informaci.

³Meióza (meiotické dělení) je proces dělení buňky probíhajícího po dvou po sobě jdoucích děleních. Výsledkem jsou buňky s haploidním (pouze jeden chromosom z každého páru) počtem chromosomů. Meiotickým dělením vznikají pohlavní buňky a meióza náhodně rozděluje mateřské a otcovské chromosomy do pohlavních buněk a tím umožňuje genetickou variabilitu.

leukémie, Alzheimerova choroba atd.

Downův syndrom se nedá vyléčit, lze však léčit některé doprovodné nemoci. Vývoj dětí s Downovým syndromem probíhá ve své podstatě téměř normálně, ale jeho proces je daleko zdlouhavější a pomalejší než u zdravých dětí, a je výsledkem intenzivního cvičení a cíleného výchovného vedení. Jejich výchova vyžaduje specifický přístup a láskyplné chování vlastních rodičů.



Obrázek 2.1: Trisomie chromosomu 21 [21]

2.3. Prenatální diagnostika

Prenatální diagnostika je soubor metod a postupů využívaných k diagnostice vrozených vad u nenarozených dětí. Jejím cílem je získat co nejpřesnější informace o aktuálním stavu a vývoji nenarozeného dítěte. A tím zajistit co nejvyšší pravděpodobnost narození zdravého dítěte. Prenatální diagnostika se neustále zdokonaluje vzrůstající odborností medicínských technologií. Problém spočívá v tom, že i přes veškeré její metody nelze nikdy na 100 % zaručit, že se dítě narodí zcela zdravé a některé z metod mohou přinést falešné pozitivní nebo negativní výsledky. Proto se názory na prenatální diagnostiku různí a některé ženy ji odmítají. Je otázkou, nakolik mají matky v České republice možnost se rozhodnout o tom, jestli chtějí podstoupit testy pre-

natální diagnostiky. V naší zemi se screening⁴ vývojových vad provádí automaticky a lékaři vlastně vůbec nepočítají s možností, že někteří rodiče si nepřipouštějí možnost usmrcení svého dítěte, i když by se mělo narodit postižené. V České republice je však základním principem dobrovolnost. Tudíž pokud rodiče z jakéhokoliv důvodu odmítají některá vyšetření, jsou lékaři povinni jejich rozhodnutí respektovat.

Na základě výsledků prenatalní diagnostiky je matka (rodiče dítěte) informována o stavu svého plodu. Pokud jí je diagnostikováno narození postiženého dítěte, žena může zvážit a následně uskutečnit ukončení svého těhotenství. V České republice je možné v souladu se současnou právní normou uměle ukončit těhotenství z genetických důvodů, a to až do 24. týdne gravidity (Zákon 66/1986 Sb. – O umělém přerušování těhotenství). V případě, kdy je zjištěno postižení, které je neslučitelné se životem dítěte, umožňuje dříve uvedený zákon uměle ukončit těhotenství kdykoliv, i po 24. týdnu gravidity.⁵

Je třeba si uvědomit, že mnoho těhotenství postižených např. Downovým syndromem končí spontánním potratem. Potratovost je různá v každém věku matky. Dále je třeba si uvědomit, že každá metoda prenatalní diagnostiky má různou *specificitu* (procento negativního nálezu) a *senzitivitu* (procento záchytu sledované nemoci). To znamená, že se jedná pouze o odhady rizika.

Postupy a metody prenatalní diagnostiky se nejčastěji dělí do dvou skupin – *neinvasivní* a *invasivní*.

2.3.1. Neinvasivní metody prenatalní diagnostiky

Neinvasivní metody jsou používané především ve screeningových programech a jsou určené pro všechny těhotné ženy. Jednotlivé metody jsou pro matku a plod téměř bezpečné a nejsou spojené s rizikem ztráty těhotenství.

⁴Screening = vyšetření, které umožňuje zjistit riziko nějaké choroby ještě před výskytem příznaků.

⁵Běžně se interrupce provádí do 12. týdne těhotenství.

Ultrazvuk

Vyšetření pomocí ultrazvukové metody patří v současné době k nejpoužívanějším vyšetřením a jedná se o běžné vyšetření v průběhu těhotenství. Provádí se v období mezi 11. – 13. týdnem těhotenství a pak v pozdějších týdnech.

Nuchální translucence (NT)

Nuchální translucence, neboli šíjové projasnění, je tekutina v kůži zátylku plodu, která je přechodně přítomna u všech plodů, ale jeho hodnota je zvýšená u plodů s chromosomálními abnormalitami, srdečními vadami a různými genetickými syndromy. Provádí se v období mezi 11. – 13. týdnem těhotenství. Nuchální translucence zachytí kolem 80 % plodů s Downovým syndromem.

Screening v 1. trimestru

Tento test je založen na počítačovém zpracování věku a počtu těhotenství matky, její váhy, ultrazvukového vyšetření tloušťky šíjového projasnění s vyšetřením krve matky. Je prováděn v období mezi 11. – 13. týdnem těhotenství. Prvotrimestrální screening má vysokou senzitivitu až kolem 98 %.

Biochemický screening

Biochemický screening se provádí ve druhém trimestru těhotenství (15. – 18. týden) a je známý jako Triple test. Je velmi jednoduchý, provádí se pouze z krve matky, ale jeho senzitivita je maximálně 67 % a téměř 15 % vychází falešně pozitivní. Donedávna se prováděl rutinně, dnes se doporučuje ho nahradit prvotrimestrálním screeningem.

2.3.2. Invazivní metody prenatální diagnostiky

Hlavním účelem těchto metod je získat vzorek tkáně plodu k cytogenetickému či molekulárně–genetickému vyšetření s cílem vyloučit u plodu chromosomální aberace či geneticky podmíněné choroby. Pro matku a plod jsou tyto metody daleko rizikovější (mohou vést ke spontánnímu potratu plodu), proto jsou prováděny jen na základě

speciální indikace. Avšak jsou více spolehlivější. Pokud je riziko postižení plodu Downovým syndromem vyšší než 1 : 300 (0,33 %), je těhotné ženě doporučováno podstoupit invazivní výkon.

Amniocentéza

Amniocentéza představuje odběr vzorku plodové vody, respektive vzorku rostoucí placenty, jehlou přes stěnu břišní pod ultrazvukovou kontrolou, které se následně testují na genetické abnormality. Plodová voda obsahuje živé buňky, a proto je možné získat informace o chromosomech nenarozeného dítěte. Vyšetření se provádí obvykle mezi 15. – 16. týdnem těhotenství. Výsledky jsou známy přibližně za dva týdny.

Odběr choriových klků (CVS)

Odběr choriových klků znamená odebrání malého množství nezralé placentární tkáně pod ultrazvukovou kontrolou. Provádí se dvěma způsoby, buď přes děložní stěnu nebo vaginální cestou. Díky tomuto vyšetření můžeme zjistit stav chromosomů plodu, jelikož placentární buňky obsahují stejnou genetickou informaci jako nenarozené dítě. Vyšetření se provádí přibližně v 11. – 14. týdnu těhotenství. Rizika u této metody jsou srovnatelná s amniocentézou a provádí se zpravidla v případě, kdy není v pořádku šíjové projasnění.

Kardocentéza

Kardocentéza znamená získání vzorku krve z pupeční žíly, jehlou přes břišní a děložní stěnu, pod ultrazvukovou kontrolou. Ze získaných krevních elementů můžeme zjistit chromosomální výbavu dítěte nebo popřípadě infekce plodu. Toto vyšetření se provádí obecně od 18. týdne těhotenství. Výsledky vyšetření jsou k dispozici velmi rychle – během 48 – 72 hodin.

2.3.3. Incidence a prevalence

Incidence a prevalence patří mezi základní ukazatele, které se používají v epidemiologii⁶. V epidemiologii se kontinuálně sleduje výskyt onemocnění či úmrtí v populaci. Je potřeba tyto jevy kvantifikovat a k tomu se využívají tzv. *ukazatele frekvence*, které vyjadřují jejich výskyt absolutními nebo relativními (poměrnými) čísly. Absolutní číslo udává celkovou četnost sledovaného jevu v populaci v daném časovém období. Relativní číslo vyjadřuje četnost sledovaného jevu vzhledem k nějakému základu jako např. velikost populace. Tento základ se často uvádí jako součet osobočasů, jenž vyjadřuje čas, po který byla daná osoba sledována. Z velké části se používá doba, po kterou byla osoba ohrožena výskytem určitého onemocnění – osobočas v riziku onemocnění.

Incidence a prevalence jsou základní ukazatelé nemocnosti, které kvantifikují výskyt onemocnění v populaci. Incidence (statistický ukazatel) kvantifikuje výskyt nově vzniklých onemocnění v populaci v daném časové intervalu (často se počítá roční incidence). Hodnota incidence se vypočte podle vzorce

$$incidence = \frac{\text{počet nových případů onemocnění v populaci v daném časovém intervalu}}{\text{součet osobočasů v riziku v populaci v daném časovém intervalu}}.$$

Prevalence reprezentuje výskyt všech existujících onemocnění v populaci ve zvoleném období a nezáleží na tom, jak dlouho onemocnění trvají. Prevalenci dělíme na dva základní typy – *okamžikovou* a *intervalovou*. Prevalence kvantifikující výskyt onemocnění v populaci v daném okamžiku je prevalence okamžiková. Okamžikem je míněno krátké období, nejčastěji se prevalence onemocnění počítá k určitému dni. Hodnota okamžikové prevalence se vypočte podle vzorce

$$okamžiková\ prevalence = \frac{\text{počet všech nemocných osob v populaci v daném okamžiku}}{\text{počet všech osob v populaci v daném okamžiku}}.$$

⁶Epidemiologie je vědní obor zabývající se zákonitostmi výskytu nemocí na populační úrovni a také faktory, které tento výskyt podmiňují nebo ovlivňují.

Prevalence kvantifikující výskyt všech existujících onemocnění ve zvoleném časovém intervalu je prevalence intervalová. Hodnota intervalové prevalence se vypočte podle vzorce

$$\textit{intervalová prevalence} = \frac{\textit{počet všech nemocných osob v populaci v daném časovém intervalu}}{\textit{součet osobočasů v riziku v populaci v daném časovém intervalu}}.$$

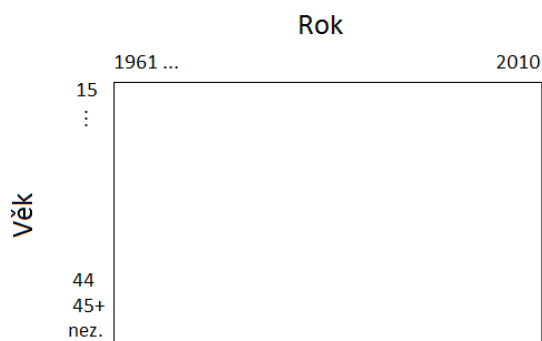
Kapitola 3

Seznámení se s daty

V této kapitole si představíme data, která jsme měli k dispozici. Pomocí grafů si ukážeme, jaké informace nám data přinášejí.

3.1. Vstupní data

Na začátku práce na této diplomové práci jsme obdrželi data v souboru Excel, která jsme měli zpracovávat. Všechna data se týkají české populace v období 1961 – 2010.



Obrázek 3.1: Struktura dat

První soubor dat představuje celkový počet narozených dětí s Downovým syndromem. Tyto údaje máme pro věkové kategorie matek 15 – 44 let, dále je tam věková kategorie matek 45⁺, což představuje matky ve věku 45 let a matky starší jak 45 let. Poslední řádek věkové kategorie, označeno jako „nezjištěný“, představují matky,

u kterých nebyl zjištěný jejich věk. Tato skupina matek je neopomenutelná, protože především v 70. a 80. letech tvoří poměrně velkou skupinu. Sloupce pak tvoří jednotlivé roky, kdy máme údaje až do roku 2010. Údaje tedy máme po jednotlivých letech (sloupce) pro věkové skupiny matek 15, 16, ..., 44, 45⁺, což s výjimkou posledního představuje kategorizaci po jednom roce věku. Na obrázku 3.1 je velice zjednodušeně znázorněna struktura dat, kdy na vodorovné ose jsou jednotlivé roky a na svislé ose je věk matek. To platí pro všechny tři soubory obdržených dat.

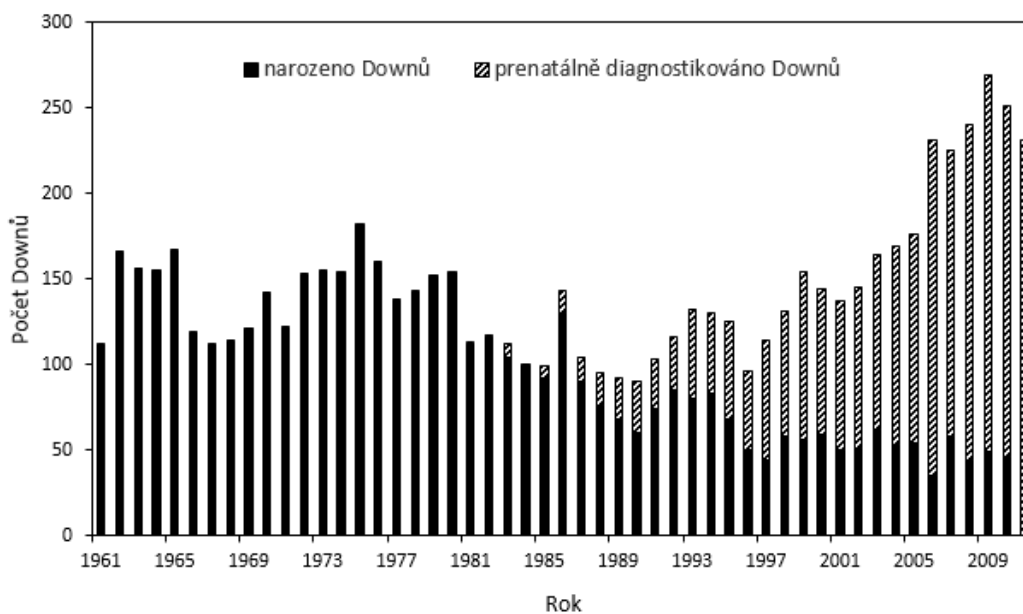
Druhý soubor dat obsahuje počty prenatalně diagnostikovaných Downů. Jako v předchozím souboru máme věkové kategorie matek a to 15 – 44 let, 45⁺ (matky ve věku 45 let a matky starší 45-ti let) a kategorii matek s nezjištěným věkem. Kategorie matek s nezjištěným věkem tvoří téměř zanedbatelnou skupinu oproti stejné kategorii matek v prvním souboru dat.

Poslední soubor dat reprezentuje celkové počty narozených dětí v České republice. Jako v předchozích dvou souborech, i zde máme věkové kategorie matek 15 – 44 let a 45⁺ (matky ve věku 45 let a matky starší 45-ti let). Rozdíl oproti předchozím dvěma souborům je, že zde není věková kategorie matek s nezjištěným věkem.

3.2. Grafy

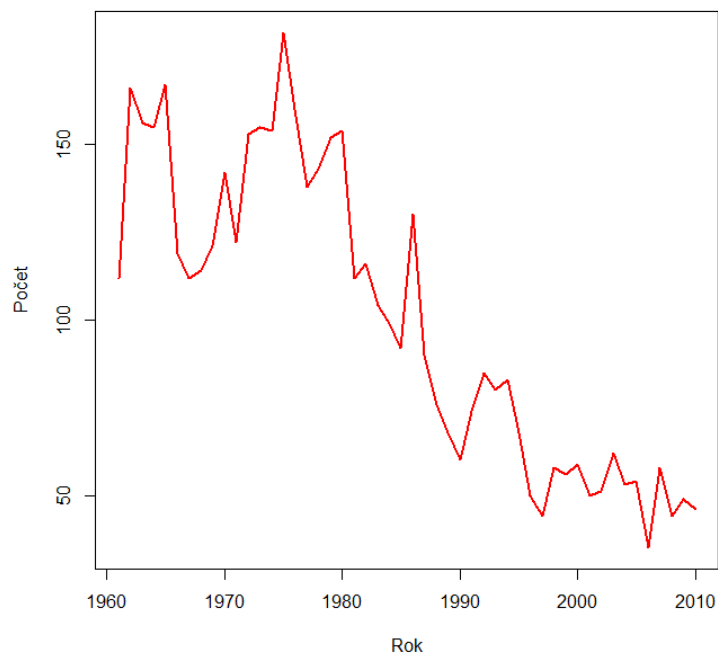
Doposud jsme se blíže seznámili s Downovým syndromem a víme, jaká budeme používat data. Nyní se pojd' me podívat, jaké informace nám obdržená data poskytují.

Jak jsme se zmínili v úvodu práce, za posledních 40 let došlo ke zlepšení prenatalní diagnostiky, která se posunula do ranějšího stádia těhotenství a která dokáže zachytit daleko více postižených plodů. Tuto skutečnost můžeme demonstrovat na obrázku 3.2, kde vidíme, jak velký přínos má prenatalní diagnostika. Zhruba od 90. let má svůj nepostradatelný význam. Černá přerušovaná barva znázorňuje, kolik bylo prenatalně diagnostikovaných Downů a černě vidíme skutečné počty narozených dětí s Downovým syndromem.



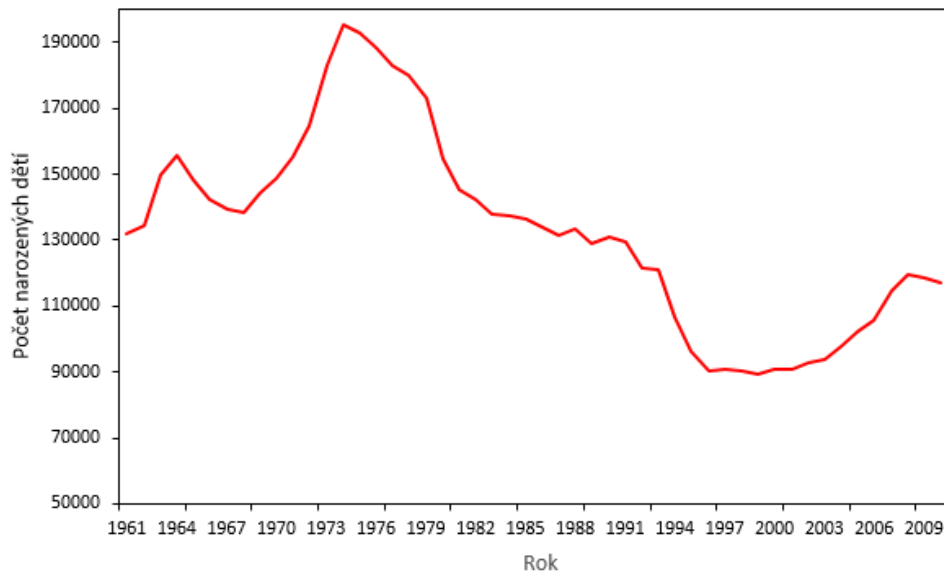
Obrázek 3.2: Rozvoj prenatální diagnostiky

Obrázek 3.3 ukazuje vývoj počtu narozených Downů. Na svislé ose jsou vynášeny skutečné hodnoty narozených Downů a na vodorovné ose vidíme vývoj v průběhu jednotlivých let. Vidíme, že od 80. let jejich počet prudce klesá, i přes vrůstající počet diagnostikovaných Downů (obrázek 3.2). To je zapříčiněno právě rozvojem prenatální diagnostiky, která zachytí více postižených těhotenství Downovým syndromem a matky pak většinou podstupují dobrovolně interrupci. Tudíž se snižuje porodnost dětí, které by byly postižené tímto syndromem.



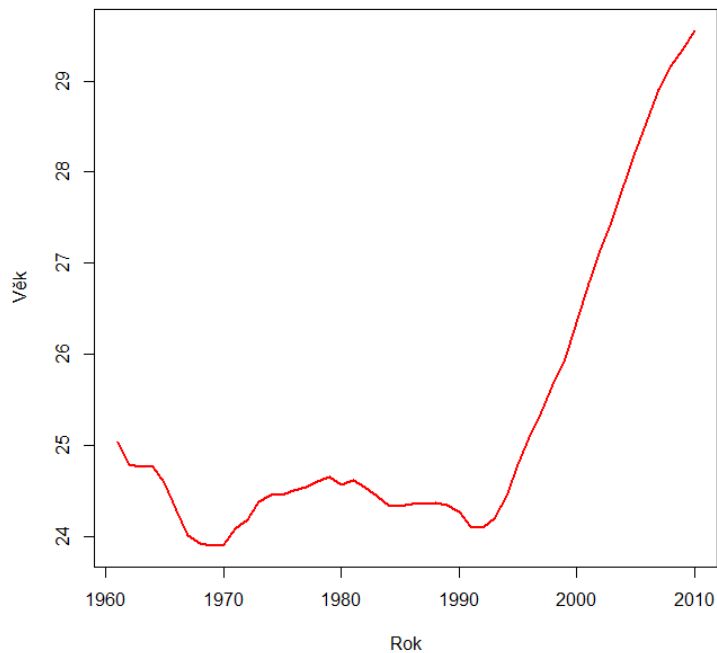
Obrázek 3.3: Celkový počet narozených Downů

Graf na obrázku 3.4 ukazuje vývoj počtu dětí narozených v České republice. Vidíme, že od 80. let má tento počet sestupnou tendenci. Snižuje se i průměrný počet dětí na jednu ženu. Tedy rodí se čím dál méně dětí, ale počet diagnostikovaných Downů naopak roste.

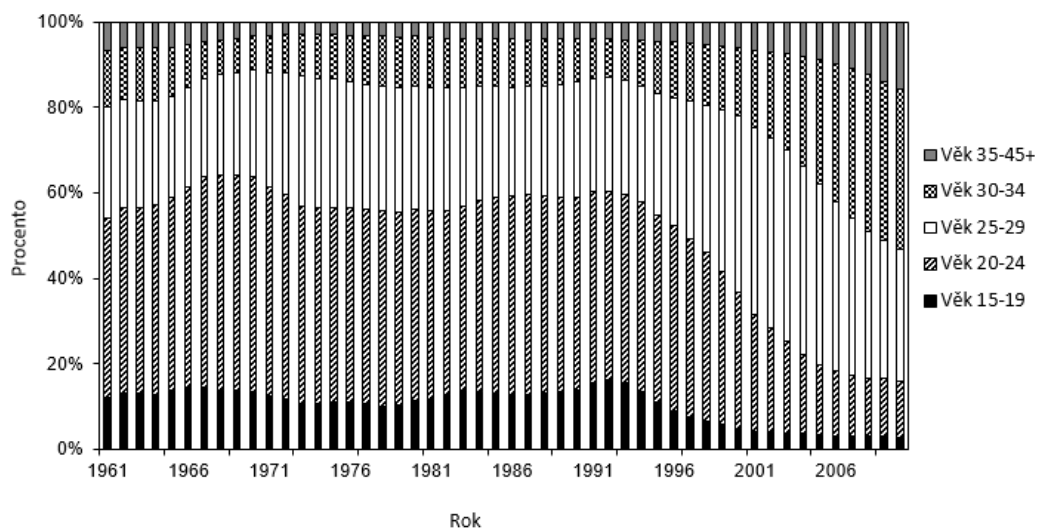


Obrázek 3.4: Počet narozených dětí v ČR (1961-2010)

Druhou podstatnou skutečností, jak už jsme se zmínili, která ovlivňuje počet diagnostikovaných Downů, je věk matek (s rostoucím věkem matky roste riziko výskytu Downova syndromu). Na grafu (obrázek 3.5) můžeme pozorovat, jak se průměrný věk matek zvyšuje. Na vodorovné ose jsou vynášeny jednotlivé roky a na svislé ose máme věk matek. Do 90. let se průměrný věk matek pohyboval pod hranicí 25 let, od 90. let můžeme pozorovat bouřlivé zvýšení tohoto průměrného věku. Tato situace je blíže zachycena na grafu (obrázek 3.6), který ukazuje procentuální složení věkových kategorií matek. Dříve (zhruba do 90. let) byla nejpočetnější skupina matek ve věku 20 – 24 let. Zastoupení této skupiny kolem roku 1994 prudce kleslo a naopak se zvyšovaly věkové skupiny matek 25 a více let.



Obrázek 3.5: Průměrný věk matek (1961-2010)



Obrázek 3.6: Věkové skupiny matek

Nyní se pojd' me podívat, jak je to s incidencí. Ve druhé kapitole bylo vysvětleno, co je incidence a k čemu se používá. V našem případě se incidence rovná poměru počtu nových případů děleno součtem počtu všech narozených jedinců (tj. počtu všech u kterých se DS mohl vyskytnout, tedy vlastně celkový počet jedinců v riziku, kde jedincem v riziku je novorozenec).

V tabulce 3.1 je vypočtená incidence Downova syndromu v České republice z našich dat. Než uvedeme jednotlivé vztahy, které jsme využili k výpočtům v tabulce 3.1, uved' me si následující značení

d_c ... celkový počet Downů,

d_n ... celkový počet narozených Downů,

$d_{pren.diag.}$... celkový počet prenatalně diagnostikovaných Downů,

I_n ... incidence narození,

I_c ... incidence celkem,

$E_{pren.diag.}$... efektivita prenatalní diagnostiky,

n ... celkový počet narozených dětí.

Jedná se vždy o údaje za jeden konkrétní rok, jelikož hodnoty v tabulce počítáme pro jednotlivé roky.

Hodnoty v tabulce jsou vypočítány podle následujících vztahů

- DS celkem

$$d_c = d_n + d_{pren.diag.}$$

- Incidence narození (přepočteno na 100 000 živě narozených dětí)

$$I_n = \frac{d_n}{n/100\,000},$$

- Incidence celkem (přepočteno na 100 000 živě narozených dětí)

$$I_c = \frac{d_c}{n/100\,000},$$

- Efektivita Pren.Diag. (v procentech)

$$E_{pren.diag.} = \frac{d_{pren.diag.}}{d_c} 100.$$

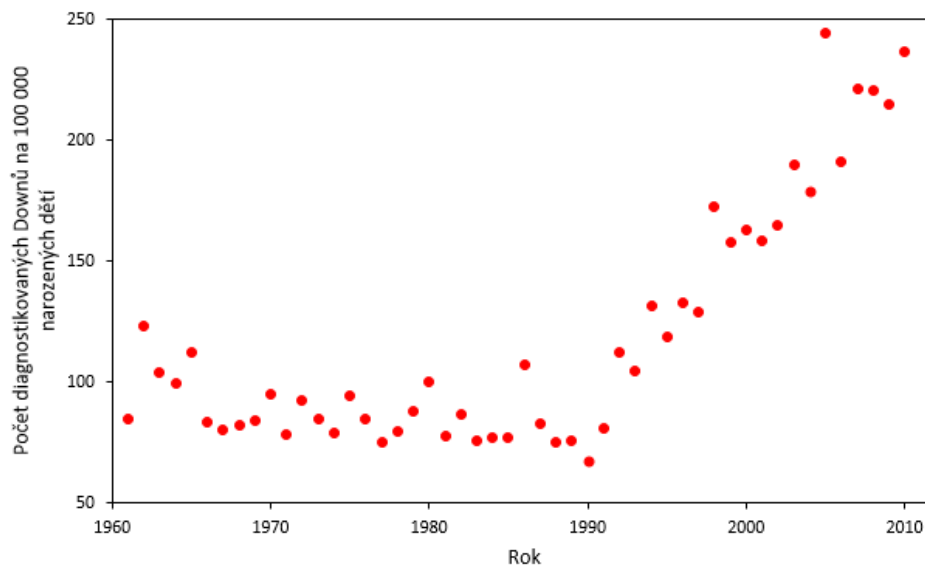
Rok	Narozeno	DS - narozeno	DS Pren.Diag.	DS celkem	Incidence narození	Incidence celkem	Efektivita Pren.Diag.
1961	132192	112	0	112	84,73	84,73	0,00
1962	134687	166	0	166	123,25	123,25	0,00
1963	150045	156	0	156	103,97	103,97	0,00
1964	155659	155	0	155	99,58	99,58	0,00
1965	148537	167	0	167	112,43	112,43	0,00
1966	142194	119	0	119	83,69	83,69	0,00
1967	139407	112	0	112	80,34	80,34	0,00
1968	138385	114	0	114	82,38	82,38	0,00
1969	144151	121	0	121	83,94	83,94	0,00
1970	148886	142	0	142	95,37	95,37	0,00
1971	155220	122	0	122	78,60	78,60	0,00
1972	164739	153	0	153	92,87	92,87	0,00
1973	182951	155	0	155	84,72	84,72	0,00
1974	195426	154	0	154	78,80	78,80	0,00
1975	192855	182	0	182	94,37	94,37	0,00
1976	188509	160	0	160	84,88	84,88	0,00
1977	182855	138	0	138	75,47	75,47	0,00
1978	179990	143	0	143	79,45	79,45	0,00
1979	173057	152	0	152	87,83	87,83	0,00
1980	154622	154	1	155	99,60	100,24	0,65
1981	145165	112	1	113	77,15	77,84	0,88
1982	142484	116	8	124	81,41	87,03	6,45
1983	138101	104	1	105	75,31	76,03	0,95
1984	137558	99	7	106	71,97	77,06	6,60
1985	136440	92	13	105	67,43	76,96	12,38
1986	133942	130	14	144	97,06	107,51	9,72
1987	131469	90	19	109	68,46	82,91	17,43
1988	133238	76	24	100	57,04	75,05	24,00
1989	128881	68	30	98	52,76	76,04	30,61
1990	131094	60	29	89	45,77	67,89	32,58
1991	129354	74	31	105	57,21	81,17	29,52
1992	121705	85	52	137	69,84	112,57	37,96
1993	121025	80	47	127	66,10	104,94	37,01
1994	106579	83	57	140	77,88	131,36	40,71

1995	96097	68	46	114	70,76	118,63	40,35
1996	90446	50	70	120	55,28	132,68	58,33
1997	90677	44	73	117	48,52	129,03	62,39
1998	90534	58	98	156	64,06	172,31	62,82
1999	89468	56	85	141	62,59	157,60	60,28
2000	89468	59	87	146	65,95	163,19	59,59
2001	90712	50	94	144	55,12	158,74	65,28
2002	92782	51	102	153	54,97	164,90	66,67
2003	93683	62	116	178	66,18	190,00	65,17
2004	97661	53	122	175	54,27	179,19	69,71
2005	102208	54	196	250	52,83	244,60	78,40
2006	105825	35	167	202	33,07	190,88	82,67
2007	114631	58	196	254	50,60	221,58	77,17
2008	119566	44	220	264	36,80	220,80	83,33
2009	118352	49	205	254	41,40	214,61	80,71
2010	117157	46	231	277	39,26	236,43	83,39

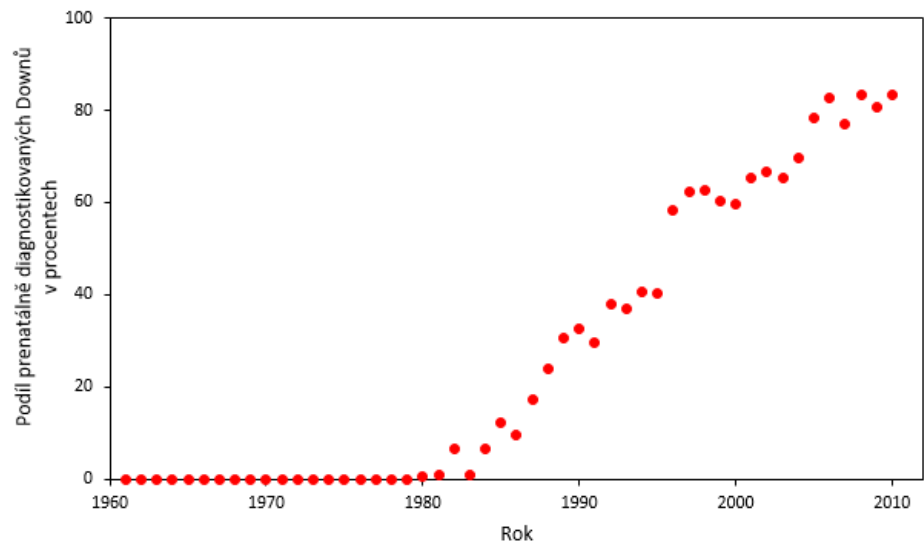
Tabulka 3.1: Incidence Downova syndromu v ČR

Vysvětlivky: **Narozeno** – počet narozených dětí v ČR; **DS narozeno** – kolik se narodilo dětí s Downovým syndromem; **DS Pren.Diag.** – kolik bylo prenatalně diagnostikováno Downů; **DS celkem** – případy Downova syndromu celkem; **Incidence narození** – incidence Downova syndromu u narozených (přepočteno na 100 000 živě narozených dětí); **Incidence celkem** – incidence Downova syndromu celková (přepočteno na 100 000 živě narozených dětí); **Efektivita Pren.Diag.** – efektivita prenatalní diagnostiky (v procentech).

Vypočtené hodnoty v tabulce 3.1 budeme demonstrovat vykreslením dvou grafů, které zobrazují následující dva obrázky. První obrázek 3.7 ukazuje celkovou incidenci Downova syndromu v České republice v období 1961 – 2010. Vidíme, že začátkem 20. let výrazně roste. Druhý obrázek 3.8 vystihuje efektivitu prenatalní diagnostiky na území České republiky. Můžeme pozorovat, že do počátku 90. let neměla žádný význam, ale pak došlo ke zlomu a s jejím vývojem se její význam začíná čím dál více zvyšovat.



Obrázek 3.7: Incidence DS (1961-2010)



Obrázek 3.8: Efektivita prenatalní diagnostiky

Kapitola 4

Riziko výskytu Downova syndromu

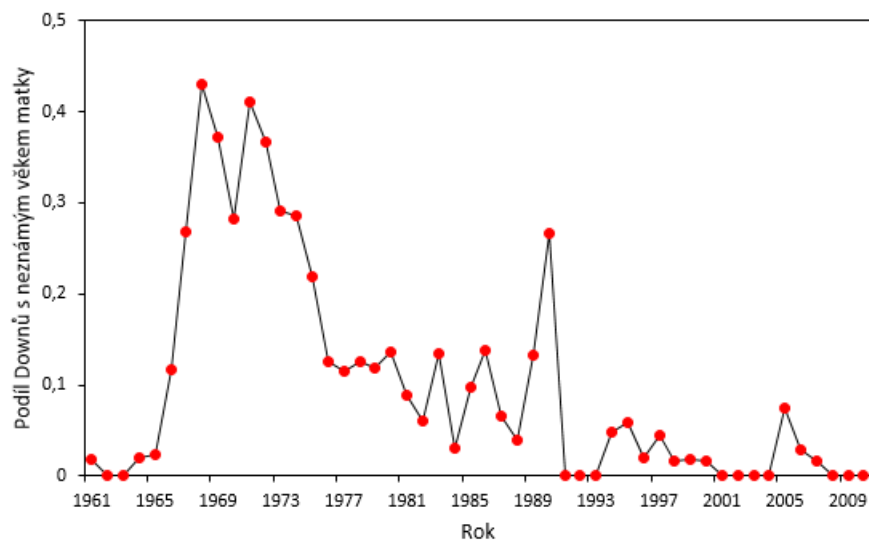
V této kapitole budeme předpokládat, že čtenář má základní povědomí z oblasti pravděpodobnosti.

Jenom stručně si připomeňme, co už všichni velmi dobře známe, a to, že riziko výskytu je popsáno pomocí pravděpodobnosti nějakého náhodného jevu (události) a může nabývat hodnot 0 až 1. Víme-li, že nějaký jev (událost) určitě nastane, pak je přirozené, že se pro tuto jistotu používá jednotková pravděpodobnost. Víme-li, že je nějaký jev (událost) nemožný, pak se mu přiřazuje nulová pravděpodobnost. Ostatním jevům (událostem), které mohou nastat, přiřazujeme pravděpodobnost mezi nulou a jedničkou.

4.1. Technický problém

Než přikročíme k výpočtu rizika výskytu Downova syndromu, narážíme na první technický problém, který kdybychom opomíjeli, tak by byly naše výsledky nepřesné. Na obrázku 3.1 je ukázka struktury dat, která máme k dispozici. Popisovali jsme si, že poslední řádek věkových kategorií matek, jsou matky s nezjištěným věkem. Tato skupina matek obsahuje mnoho hodnot, proto je nemůžeme zanedbávat. Takže potřebujeme tyto hodnoty nějak vhodně roztřídit (odhadnout jejich počty) k věkovým kategoriím matek, jejichž věk známe. Situaci zachycuje obrázek 4.1. Na vodorovné ose jsou jednotlivé roky a na svislé ose jsou podíly Downů s neznámým věkem matky. Vidíme, že zhruba roku 1966 došlo k prudkému nárůstu tohoto podílu a jeho počty do-

sahovaly obrovských hodnot. Poté došlo k rychlému poklesu a následně opět k nárůstu. V 80. letech můžeme pozorovat velký pokles a v 90. letech byla situace poměrně nestabilní. Od roku 1991 se podíly Downů s nezjištěným věkem matky pohybovaly v nízkých hodnotách. Dále budeme postupovat způsobem, který je popsán níže.



Obrázek 4.1: Podíl Downů s neznámým věkem matky

Uvažujme jeden konkrétní libovolný rok a v tomto roce označme

n_i ... celkový počet narozených dětí matkám v i -té věkové kategorii,

d_i ... počet Downů narozených matkám v i -té věkové kategorii.

Tyto údaje jsou známe. Neznámý údaj označme jako

D_i ... skutečný počet Downů narozených matkám v i -té věkové kategorii.

Chtěli bychom vypočítat odhad rizika Downova syndromu v i -té věkové kategorii

$$\frac{D_i}{n_i}.$$

Dále ještě označíme

d_{nez} ... počet Downů s chybějícím údajem o věku matky,

$d = \sum_i D_i = \sum_i d_i + d_{nez}$... celkový počet narozených Downů,

$P_i = 1 - \frac{d_i}{D_i} = \frac{D_i - d_i}{D_i}$... podíl nezjištěných, tzn. podíl Downů v i -té věkové skupině,

u nichž nemáme údaj o věku matky.

Platí

$$\frac{D_i}{n_i} = \frac{d_i}{n_i(1 - P_i)}.$$

Budeme předpokládat, že pravděpodobnost chybějícího údaje je ve všech věkových skupinách matek stejná a proto můžeme psát

$$EP_i = EP.$$

Tuto pravděpodobnost odhadneme jako

$$p := \widehat{EP}_i = \frac{d_{nez}}{d}.$$

Dále potřebujeme riziko Downova syndromu v i -té věkové skupině a to odhadneme jako

$$\frac{D_i}{n_i} = \frac{d_i}{n_i(1 - P_i)} \approx \frac{d_i}{n_i(1 - p)}.$$

Označením

$$n_i^* = n_i(1 - p)$$

dostáváme odhad rizika ve tvaru

$$\frac{d_i}{n_i^*}.$$

4.2. Výpočet rizika výskytu Downova syndromu

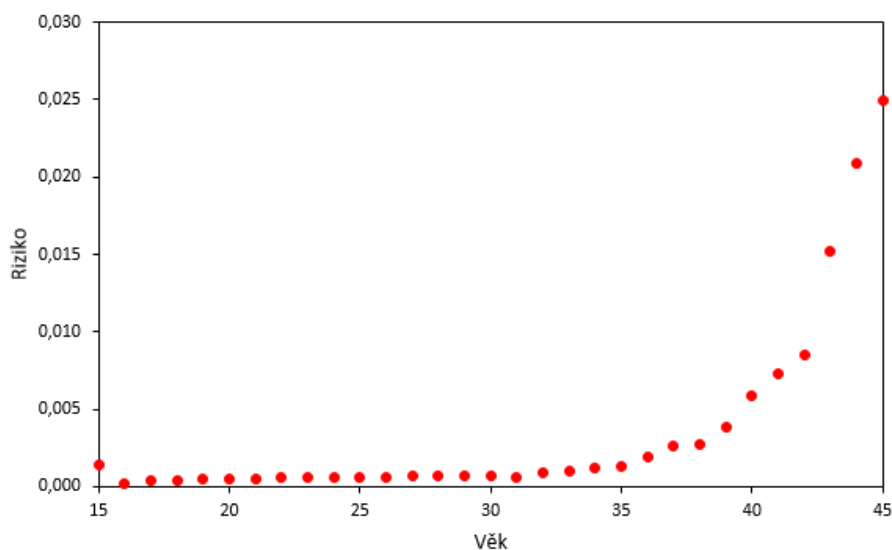
Vypočítáme riziko (pravděpodobnost) Downova syndromu pro věkové kategorie matek po úpravě o případy, kdy věk matky nebyl zjištěný, tak jak bylo popsáno v podkapitole 4.1. Pak postupujeme podle následujícího vztahu

$$p(v) = \frac{\text{počet Downů } (v)}{\text{počet všech } (v)}, \quad (4.1)$$

kde

$p(v)$... pravděpodobnost Downova syndromu pro jednotlivé věkové kategorie matek,
počet Downů (v) ... počet narozených Downů pro jednotlivé věkové kategorie matek,
počet všech (v) ... celkový počet narozených dětí pro jednotlivé věkové kategorie matek.

Vypočítané hodnoty vyneseme do grafu (obrázek 4.2), kde na svislou osu vynášíme hodnoty rizika a na vodorovnou osu vynášíme věk matky. Vidíme, že riziko s věkem matky výrazně roste. Jelikož je graf (obrázek 4.2) špatně čitelný, je vhodnější pracovat s empirickým logitem rizika.



Obrázek 4.2: Riziko výskytu Downova syndromu

Pravděpodobnost, jak již bylo řečeno, nabývá hodnot mezi 0 a 1. Jiným používaným způsobem zápisu pravděpodobnosti je zápis pomocí šancí jevu (*odds*). Šance je definována jako pravděpodobnost, že jev nastal, dělená pravděpodobností, že jev nenastal. Šance nemá žádnou pevnou maximální hodnotu, ale její minimálně hodnota je rovna 0. [14]

$$\text{odds}(v) = \frac{p(v)}{1 - p(v)}, \quad (4.2)$$

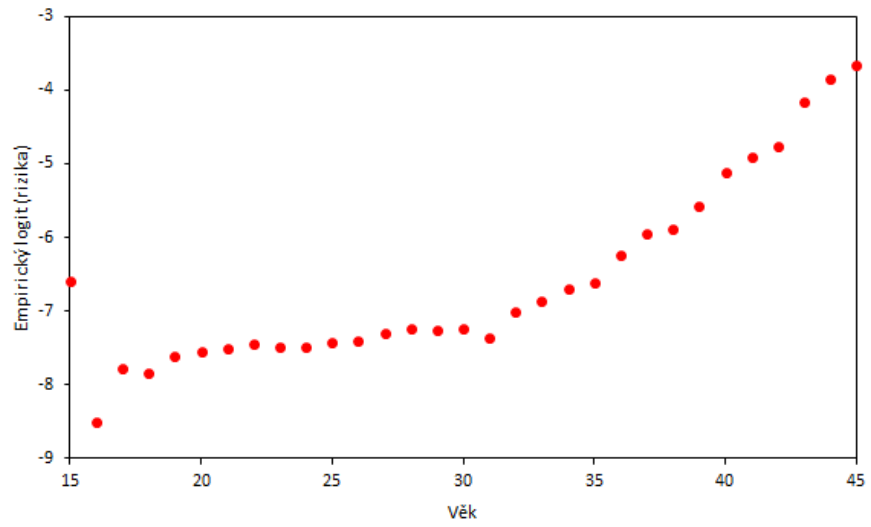
$p(v)$... pravděpodobnost Downova syndromu pro jednotlivé věkové kategorie matek.

Šance tedy leží v intervalu $(0, \infty)$. Chceme vzájemně jednoznačně transformovat interval $(0, \infty)$ na interval $(-\infty, \infty)$. K tomuto účelu se používá funkce přirozeného logaritmu, čímž zavedeme logitovou funkci [14]

$$\text{logit}(p(v)) = \ln\left(\frac{p(v)}{1 - p(v)}\right), \quad (4.3)$$

$p(v)$... pravděpodobnost Downova syndromu pro jednotlivé věkové kategorie matek.

Odhadnuté hodnoty logitu, podle vztahu 4.3, vyneseme do grafu (obrázek 4.3), kde na svislé ose je logit rizika a na vodorovné ose věk matky. Vidíme, že začátek je poměrně „divoký“, to si vysvětlujeme především velkou variabilitou odhadů při nižších počtech narozených dětí v těchto věkových kategoriích matek. Pak můžeme pozorovat, že křivka má tendenci velice nepatrně růst, kdy ve věku 31 let mírně klesá. Od věku 31 let křivka začíná růst prudčeji.



Obrázek 4.3: Logit rizika výskytu Downova syndromu

Kapitola 5

Model závislosti rizika Downova syndromu na věku matky

V zahraničních referenčních studiích se můžeme setkat s různými modely závislosti rizika Downova syndromu na věku matky. Na obrázku 5.1 máme 6 různých modelů od 5 skupin autorů z různých let. Každý z modelů modeluje tuto závislost trochu odlišným způsobem. Modely se především liší v tom, jak je modelován konec křivky. Modely od autorů Cuckle et al. (1987), Hecht and Hook (1994) a Bray et al. (1998) předpokládají, že křivka se zvyšujícím se věkem matky bude stále rostoucí. Huether et al. (1998) uvažují, že křivka bude též rostoucí, ale už si můžeme povšimnout, že kolem 50. roku věku ženy se začíná mírně zplošťovat. Od Hetche a Hooka tu máme ještě jeden model a to z roku 1996, kde oproti jejich předešlému modelu ukazují, že konec křivky by měl zhruba od věku ženy 49 let klesat. Poslední model je od autorů Morrise, Muttona a Albermanna (2002), kteří se domnívají, že konec křivky od 45. roku života ženy bude víceméně stabilní. Tyto modely jsou pro nás možným vodítkem či inspirací, jak bychom k modelování závislosti rizika DS na věku matky mohli přistupovat. Je otázkou, který ze zmíněných modelů by se dal použít a jestli vůbec je některý z nich použitelný na námi analyzovaná data. Na obrázku 5.2 jsou všechny výše zmíněné modely shrnuty v jednom grafu a pod grafem můžeme vidět jednotlivé rovnice použité při modelování závislosti rizika Downova syndromu na věku matky. Čerpáno z [11], [12].

Budeme se snažit námi analyzovanými daty proložit křivku a tím vytvořit co nej-

výstižnější model. Než se dostaneme k samotnému modelování, popišme si metody, se kterými budeme následně pracovat.

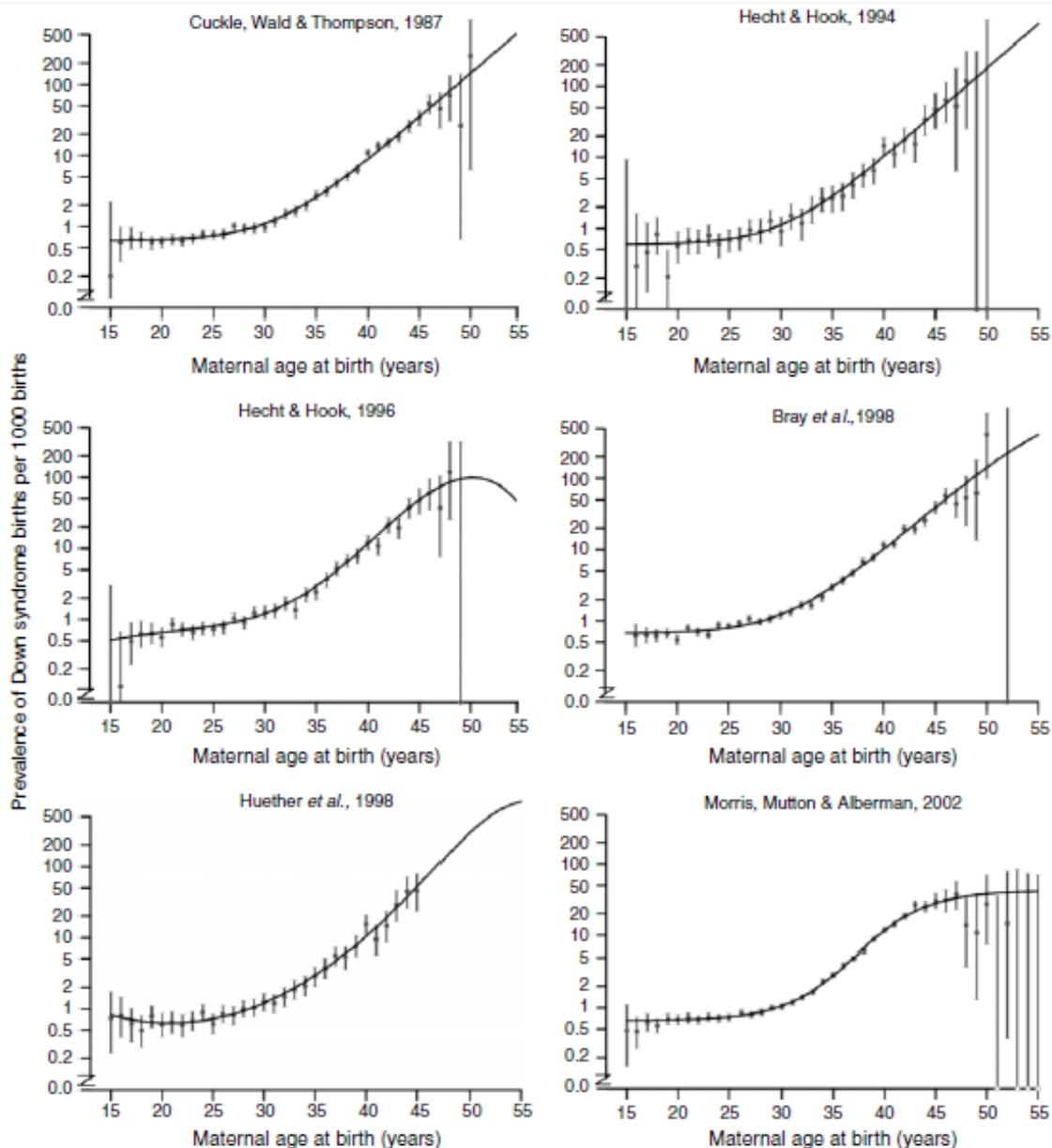


Figure 1—Down syndrome live birth rates by single year of maternal age with 95% confidence intervals and the predicted prevalence of Down syndrome live births

Obrázek 5.1: Modely závislosti rizika DS na věku matky I. [11]

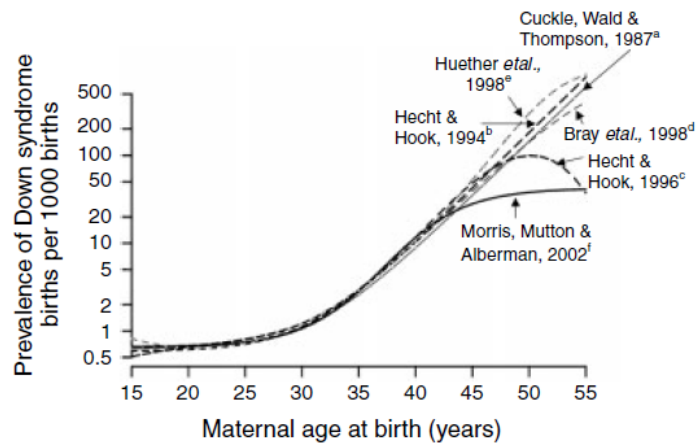


Figure 2—Predicted prevalence of Down syndrome live birth rates. Fitted models are ^aRisk = $0.000627 + \exp(-16.2395 + 0.286 * \text{mage})$; ^bRisk = $0.0005871 + \exp(-16.2882036 + 0.2915656 * \text{mage})$; ^cRisk = $0.0003533 + \exp(-21.9647716 + 1.9777067 * \text{mage} - 0.1060169 * \text{mage}^2 + 0.0024915 * \text{mage}^3 - 0.0000201 * \text{mage}^4)$; ^dRisk = $0.0006589 + (1 - 0.0006589) / \{1 + \exp[-(-16.024236 + 0.2844376 * \text{mage})]\}$; ^eRisk = $1 / \{1 + \exp[-(-4.061283 - 0.317749 * \text{mage} - 0.007638 * \text{mage}^2)]\}$; ^fRisk = $1 / \{1 + \exp(7.330 - 4.211 / [1 + \exp[-0.282 * (\text{mage} - 37.23)])]\}$; where mage = maternal age at birth (years)

Obrázek 5.2: Modely závislosti rizika DS na věku matky II. [11]

5.1. Logistická regrese

Metoda logistické regrese byla navržena v 60. letech 20. století jako alternativní metoda k metodě lineární regrese. Používá se pro případ, kdy vysvětlovaná (závisle) proměnná je binární (nabývá pouze dvou hodnot a to 0 nebo 1). V minulosti se nejvíce využívala v medicíně a v epidemiologii, kdy vysvětlovaná (závisle) proměnná představuje přítomnost či nepřítomnost nějaké choroby. Logistická regrese se od lineární regrese liší tím, že predikuje pravděpodobnost toho, jestli se nějaká událost stane nebo nestane. K vytvoření vazební podmínky mezi touto pravděpodobností a lineárním prediktorem tvořeným nezávislými proměnnými se používá logitová transformace (4.3,

vychází z poměru šancí (4.2)), která vede k sigmoidálnímu vztahu mezi závisle proměnnou a vektorem nezávisle proměnných. Logistická regrese používá kategorickou vysvětlovanou (závisle) proměnnou, kdežto u lineární regrese je vysvětlovaná (závisle) proměnná spojitá.

Rozlišujeme tři typy logistické regrese, podle typu vysvětlované proměnné

- *Binární logistická regrese*, která má vysvětlovanou (závisle) proměnnou binární. Ta nabývá pouze dvou možných hodnot, například absence a přítomnost jevu.
- *Ordinální logistická regrese*, která má ordinální (závisle) proměnnou nabývající tří a více možných stavů, mezi nimiž existuje přirozené uspořádání, například typ odpovědi v dotazníku s možnými odpověďmi – vůbec ne, spíše ne, spíše ano, určitě ano.
- *Nominální logistická regrese*, která má nominální (závisle) proměnnou obsahující více než dvě úrovně různých stavů, jež mohou definovat pouze odlišnost, například barva očí.

Nezávislé (vysvětlující) proměnné označme $\mathbf{X} = (X_1, \dots, X_m)'$ a závislou (vysvětlovanou) proměnnou Y . Transformovanou pravděpodobnost můžeme modelovat obdobně jako u lineární regrese

$$\ln\left(\frac{p(Y)}{1-p(Y)}\right) = \alpha + \beta_1 X_1 + \dots + \beta_m X_m, \quad (5.1)$$

kde α a $\beta = (\beta_1, \dots, \beta_m)'$ jsou parametry.

Odtud vyjádříme pravděpodobnost

$$p(Y) = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \dots + \beta_m X_m)}}. \quad (5.2)$$

Při označení α , $\beta = (\beta_1, \dots, \beta_m)'$ a $\mathbf{X} = (X_1, \dots, X_m)'$ můžeme pravděpodobnost přepsat jako

$$p(Y) = \frac{1}{1 + e^{-(\alpha + \mathbf{X}'\beta)}}. \quad (5.3)$$

Pravděpodobnost můžeme přepsat pomocí podmíněné pravděpodobnosti

$$p(Y = 1|X_1, \dots, X_m) = \frac{1}{1 + e^{-(\alpha + \mathbf{X}'\beta)}}. \quad (5.4)$$

Následující definice je definicí logistického modelu.

Definice 5.1.1. Logistický model pravděpodobnosti v základním tvaru definujeme jako

$$p(Y|X_1, \dots, X_m) = p(Y|\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \mathbf{X}'\beta)}}, \quad (5.5)$$

kde α , $\beta = (\beta_1, \dots, \beta_m)'$ jsou neznámé parametry, $\mathbf{X} = (X_1, \dots, X_m)'$ jsou nezávislé proměnné a Y je alternativně rozdělená závisle proměnná.

5.1.1. Metoda maximální věrohodnosti

U modelu logistické regrese se k odhadům koeficientů modelu nejčastěji používá *metoda maximální věrohodnosti* (*maximum likelihood estimation*, zkratka MLE). Základním principem metody maximální věrohodnosti je najít odhad parametru θ , který maximalizuje pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení pravděpodobnosti. Jinak řečeno, snažíme se najít takovou hodnotu parametru θ , pro niž je pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení pravděpodobnosti, maximální.

Nechť $\mathbf{X} = (X_1, \dots, X_n)'$ je náhodný výběr z diskrétního rozdělení s pravděpodobnostní funkcí $p(\mathbf{x}, \Theta)$, resp. ze spojitého rozdělení s hustotou $f(\mathbf{x}, \Theta)$, kde $\mathbf{x} \in \mathbb{R}^n$, $\Theta = (\theta_1, \dots, \theta_m)' \in \Omega$ je vektorem neznámých parametrů. Tento vektor neznámých parametrů může nabývat pouze hodnot z nějakého parametrického prostoru $\Omega \subset \mathbb{R}^m$.

Pro jednoduchost budeme uvažovat jednorozměrný parametr θ . Pro každé pevné $\mathbf{x} \in \mathbb{R}^n$ můžeme pravděpodobnostní funkci $p(\mathbf{x}, \theta)$, resp. funkci hustoty $f(\mathbf{x}, \theta)$, chápat jako funkce proměnné θ . Tuto funkci označíme jako $L(\theta)$ (z ang. *likelihood* neboli *věrohodnost*) a označíme ji jako *věrohodnostní funkce*. Pro diskrétní vzájemně nezávislé, stejně rozdělené náhodné veličiny s pravděpodobnostní funkcí $p(\mathbf{x}, \theta)$ má věrohodnostní funkce tvar

$$L(\theta) = \prod_{i=1}^n p(x_i, \theta). \quad (5.6)$$

Pro spojité vzájemně nezávislé, stejně rozdělené náhodné veličiny s hustotou pravděpodobnosti $f(\mathbf{x}, \theta)$ má věrohodnostní funkce tvar

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta). \quad (5.7)$$

Jestliže existuje hodnota $\hat{\theta} \in \Omega$ taková, že pro všechny možné hodnoty parametru $\theta \in \Omega$ platí

$$L(\mathbf{x}, \theta) \leq L(\mathbf{x}, \hat{\theta}), \quad (5.8)$$

pak nazveme odhad $\hat{\theta}$ *maximálně věrohodným odhadem* neznámého parametru θ . Alternativní formulace je

$$\hat{\theta} = \arg \max_{\theta \in \Omega} L(\theta). \quad (5.9)$$

Často je výhodnější místo s věrohodnostní funkcí pracovat s jejím přirozeným logaritmem (*logaritmická věrohodnostní funkce*)

$$l(\theta) = \ln L(\theta). \quad (5.10)$$

Uvědomíme-li si, že logaritmus součinu je rovný součtu logaritmů, můžeme maximálně věrohodný odhad $\hat{\theta}$ parametru θ zpravidla stanovit řešením tzv. *věrohodnostních rovnic*

$$\frac{\partial \ln L(\theta)}{\partial \theta_j} = 0, \quad j = 1, \dots, m. \quad (5.11)$$

Čerpáno zejména z [9], [10].

5.2. Joinpoint regrese

Joinpoint regrese (*segmented regression, piecewise regression, broken line regression*) je model, ve kterém je závislost vysvětlované veličiny popsána po částech lineární spojitou funkcí. Tato funkce mění svou směrnici v několika obecně neznámých bodech zvratu (*joinpoints*). Cílem je tedy nalézt právě tyto body zvratu. Výhodou joinpoint regrese je především snadná interpretace parametrů, která nám umožňuje jednoduše

popsat změny ve sledované závislosti. Joinpoint regresi můžeme využívat i v případech složitějších modelů, jako jsou zobecněné lineární modely (např. logistická regrese, loglineární model) a další. Poslední dobou se tato regrese používá velmi hojně v medicíně.

Pro odhad po částech lineární spojité regresní funkce můžeme použít program R a v něm využít knihovnu `segmented`. Tato knihovna umožňuje odhadovat joinpoint model pro zobecněné lineární modely a dovoluje pracovat s více vysvětlujícími proměnnými. Avšak tato knihovna se nezabývá odhadem počtu bodů zvratu.

Odhad modelu je proveden speciálním iteračním postupem. Chceme-li odhadnout joinpoint model, který má zvrát v bodě ψ a ve kterém se směrnice změní z β_1 na $\beta_2 = \beta_1 + \delta_1$, tak model je ve tvaru ¹

$$E(Y) = \beta_0 + \beta_1 x + \delta_1 (x - \psi)^+, \quad (5.12)$$

kde β_0 , β_1 a δ_1 jsou neznámé regresní koeficienty, které se odhadují metodou maximální věrohodnosti, ψ je bod zvratu a znaménko plus značí kladnou část. Je-li ψ_0 počáteční odhad bodu zvratu ψ , pak můžeme model 5.12 odhadnout iterativním odhadováním následujícího lineárního modelu

$$E(Y) = \beta_0 + \beta_1 x + \delta_1 (x - \psi_0) + \gamma I(x > \psi_0), \quad (5.13)$$

kde $I(\cdot)$ je identifikátor a γ je parametr. Parametr γ měří nespojitost v bodě zvratu a pomocí něj je přepočítáván odhad ψ_0 . Je-li algoritmus konvergentní, pak by měla být výsledná regresní funkce spojitá, tj $\hat{\gamma} \approx 0$.

Nevýhodou je, že věrohodnostní funkce v joinpoint modelu nemusí být konkávní, to znamená, že algoritmus nemusí nalézt globální minimum. Další nevýhodou je, že navržený algoritmus v každém kroku pouze aproximuje skutečný model 5.12, což může způsobovat problémy při odhadování. V praxi je proto dobré spustit program pro několik různých počátečních hodnot parametru ψ . Jestliže dostaneme různé hodnoty odhadu bodu zvratu, můžeme program požádat o výpis věrohodnosti, která od-

¹U zobecněných lineárních modelů je model ve tvaru

$$g(EY) = \beta_0 + \beta_1 x + \delta_1 (x - \psi)^+,$$

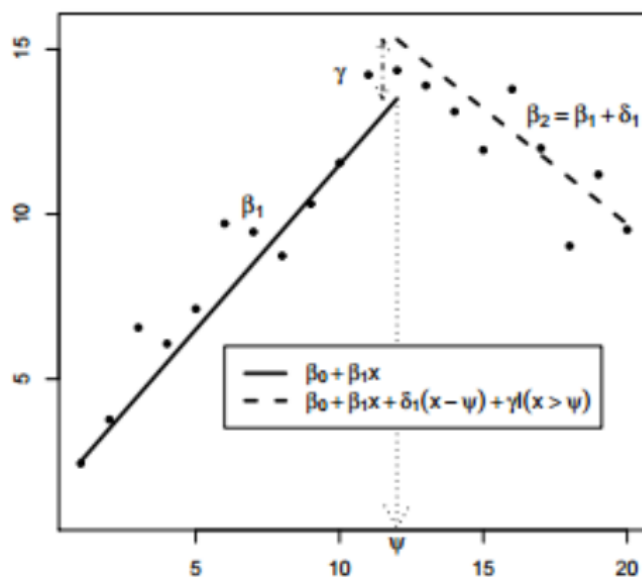
kde g je nějaká linková funkce (v našem případě to je logit), β_0 , β_1 a δ_1 jsou neznámé regresní koeficienty, ψ je bod zvratu a znaménko plus značí kladnou část.

povídá jednotlivým modelům a „ručně“ vybrat ten nejvhodnější. Doporučuje se volit počáteční hodnotu bodu zvratu na základě posouzení grafického znázornění námi sledované závislosti.

Problém, který může nastat, je, že algoritmus nekonverguje, na což nás většinou upozorní program, který vypíše varovnou hlášku, která říká, že byl dosažen maximální počet iterací. Jako řešení by se mohlo zdát, že stačí navýšit počet povolených iterací, ale to bohužel není řešení, neboť tato situace často nastává v případě, kdy minimalizovaná funkce alternuje mezi dvěma hodnotami.

Jestliže konvergence není dosažena automaticky nebo v případě, kdy obdržíme různé výsledky pro různá počáteční nastavení parametru ψ , může být parametrizace joinpoint modelem pro daná data diskutabilní. Můžeme pak provést například test přítomnosti bodu zvratu.

Čerpáno z [14], [19].



Obrázek 5.3: Model, prostřednictvím něhož je iterativně odhadnuta joinpoint regrese pomocí funkce segmented. [19]

5.3. Model

V této podkapitole se již dostáváme k hlavní části této diplomové práce a to k vytvoření modelu, který bude korespondovat s našimi analyzovanými daty. Chtěli jsme vytvořit model závislosti rizika výskytu Downova syndromu na věku matky. Byl to úkol nelehký, neboť na počátku práce jsme neměli představu, kterou metodu použít. Jedinými indiciemi nám byly (již zmiňované) zahraniční referenční studie (obrázek 5.1).

Jako nejvhodnější se jevílo použít metodu logistické regrese a parametry odhadnout pomocí metody maximální věrohodnosti. Další postup, co jsme si vyzkoušeli, byla joinpoint regrese v programu R. Nyní se podíváme podrobněji na postupy řešení a získané výsledky.

5.3.1. Řešení pomocí logistické regrese

Naše analyzovaná data mají binomické rozdělení pravděpodobnosti, pokud se na ně díváme tak, že událost nastala nebo nenastala (událost čili výskyt DS). Logistická regrese modeluje závislost pravděpodobnosti, že událost nastane, na věku matky. Tedy modelujeme binomickou proměnnou. Výskyt DS u jednotlivých novorozenců je alternativně rozdělená náhodná veličina a jednotlivé výskyty považujeme za nezávislé, tedy jejich počet v daném roce v dané věkové kategorii je binomicky rozdělená náhodná veličina. Pro libovolnou věkovou kategorii považujeme počty Downů v jednotlivých letech za nezávislé binomické veličiny.

Riziko výskytu Downova syndromu popisuje logistická regrese na oboru hodnot 0 (žádné riziko) až 1 (jistá událost). Definičním oborem je věk matky, který je kladný. Praktický význam má sledovat pouze věk od 15 let (jak už jsme si uvedli, máme data pro věkové kategorie matek 15 až 44 let a věkovou kategorii matek 45⁺, kde jsou matky starší než 45 let). Model se skládá ze dvou křivek – do 30. roku života matky z jedné křivky a od 31. roku života matky z druhé křivky. Zlomovým bodem je tedy věk matky 31 let. Tuto hodnotu jsme našli tak, že jsme zkoušeli dosazovat různé věky v do námi navrženého modelu. Vypočítali jsme hodnoty parametrů a zjišťovali jsme,

pro kterou hodnotu dělicího bodu (věku matky) dostaneme největší hodnoty maxima věrohodností funkce.

Nejdříve jsme modelovali transformovanou pravděpodobnost (riziko) pomocí vztahů 4.3, 5.1

$$\text{logit}(p_v) = \begin{cases} a + bv & \text{pro } v < 31, \\ a + bv + c(v - 31) & \text{pro } v \geq 31, \end{cases} \quad (5.14)$$

kde v je věk matky a a, b, c jsou parametry modelu.

Model je pak ve tvaru (pomocí vztahu 5.2)

$$p_v \approx \begin{cases} \frac{1}{1 + e^{-(a+bv)}} & \text{pro } v < 31, \\ \frac{1}{1 + e^{-(a+bv+c(v-31))}} & \text{pro } v \geq 31. \end{cases} \quad (5.15)$$

Přičemž v je věk matky a a, b, c jsou parametry modelu, které budeme odhadovat metodou maximální věrohodnosti.

Označme vektor neznámých parametrů jako $\theta = (a, b, c)$. Nejprve vyjádříme funkci věrohodnosti, která je pro $y_v, v = 15, \dots, 45^+$ a libovolné hodnoty parametru θ ve tvaru

$$L(\theta) = \prod_v f(y_v, \theta). \quad (5.16)$$

y_v je riziko Downova syndromu pro věkové kategorie matek upravené o případy, kdy věk matky nebyl zjištěný, vyjádřené jako (4.1 – pouze s jiným značením)

$$y_v = \frac{\text{počet Downů } (v)}{\text{počet všech } (v)},$$

kde

počet Downů (v) ... počet narozených Downů pro jednotlivé věkové kategorie matek,
počet všech (v) ... celkový počet narozených dětí pro jednotlivé věkové kategorie matek.

Užitím předchozího dostáváme pro model 5.15 věrohodnostní funkci ve tvaru

$$L(\theta) = \prod_v (p_v^{y_v} (1 - p_v)^{(1 - y_v)}). \quad (5.17)$$

Z výpočetních důvodů použijeme logaritmus věrohodnostní funkce, což můžeme, neboť logaritmus nějaké funkce má maximum ve stejném bodě jako samotná funkce. Po zlogaritmování dostáváme

$$\ln L(\theta) = \sum_v (\ln(p_v) y_v + \ln(1 - p_v) (1 - y_v)). \quad (5.18)$$

Chceme najít takovou hodnotu vektoru parametru θ , která bude maximalizovat věrohodnostní funkci. Hledáme tedy odhad $\hat{\theta}$

$$\hat{\theta} = \arg \max_{\theta} \sum_v \ln f(y_v, \theta). \quad (5.19)$$

Odhad $\hat{\theta}$ vektoru parametru θ podle vztahu 5.19 ještě není úplně správný. Jak už jsme si říkali, máme věkové kategorie matek od 15 let až do věku 44 a věkovou kategorii matek 45⁺, která zahrnuje matky starší jak 45 let. Pro tyto věkové kategorie matek máme údaje za období 1961 – 2010. Na tomto místě si však musíme uvědomit, že každá věková kategorie matek přispívá různou mírou do našeho modelu. Některé kategorie jsou daleko početnější než jiné. Musíme tedy stanovit váhu každé věkové kategorie matek a tuto váhu zahrnout do výpočtu odhadu neznámého vektoru parametru θ .

Tuto váhu určíme ze vztahu

$$\frac{\sum_v w(v)}{m} = 1,$$

kde

m ... počet věkových kategorií,

$\sum_v w(v)$... součet celkově narozených Downů pro jednotlivé věkové kategorie matek (počty narozených Downů upravené o případy, kdy věk matky nebyl zjištěný).

Když už známe váhy $w(v)$, kterými přispívají jednotlivé věkové kategorie matek, upravíme vztah 5.19 o tyto váhy a dostáváme

$$\hat{\theta} = \arg \max_{\theta} \sum_v \ln f(y_v, \theta) w_v. \quad (5.20)$$

Metodou maximální věrohodnosti jsme vypočítali následující odhad parametrů (zaokrouhлено na pět desetinných míst)

$$a = -8,02823,$$

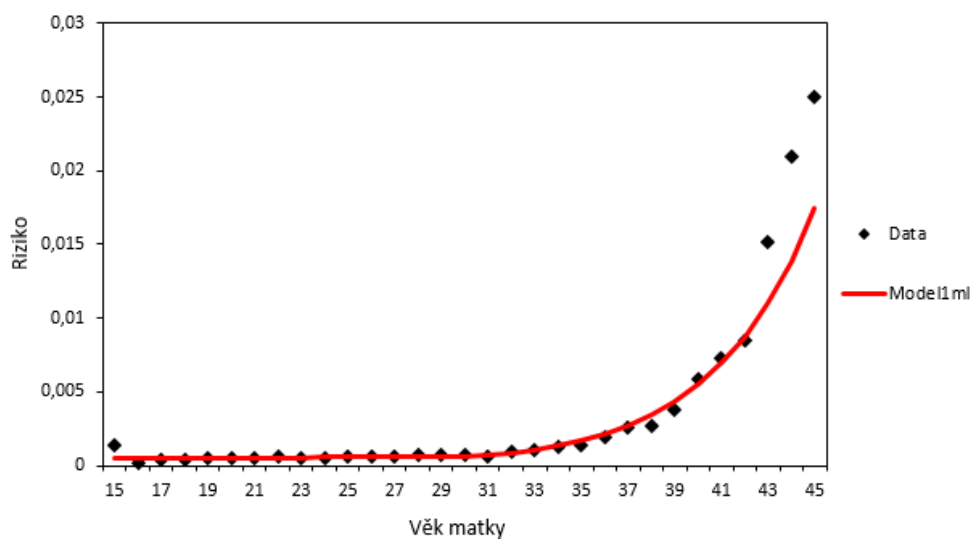
$$b = 0,02372,$$

$$c = 0,20913$$

a logaritmus věrohodnosti činí $-0,18494$.

Hodnoty parametrů dosadíme do našeho modelu 5.15 a dostáváme tedy model s odhadnutými parametry

$$p_v \approx \begin{cases} \frac{1}{1 + e^{-(-8,02823 + 0,02372v)}} & \text{pro } v < 31, \\ \frac{1}{1 + e^{-(-8,02823 + 0,02372v + 0,20913(v-31))}} & \text{pro } v \geq 31. \end{cases} \quad (5.21)$$



Obrázek 5.4: Model závislosti rizika DS na věku matky

Na obrázku 5.4 vidíme grafický výstup námi odhadnutého modelu. Můžeme si povšimnout, že výsledek vypadá celkem obstojně.

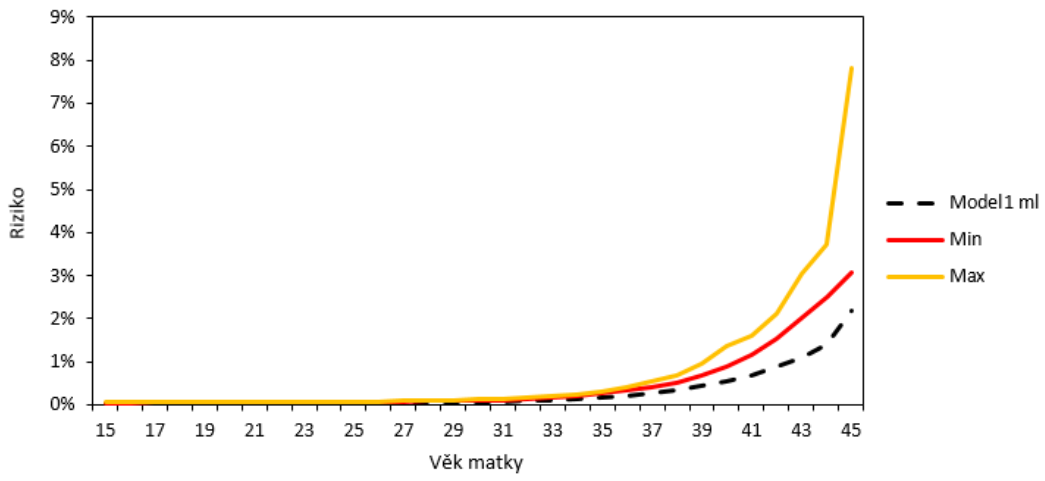
Jako velice zajímavé se jevílo, že bychom mohli porovnat riziko, které jsme vyjádřili modelem 5.15, s riziky, které jsou uvedené v zahraničních referenčních studiích (obrázek 5.5). Na obrázku 5.5 vidíme hodnoty rizika výskytu Downova syndromu přepočtené na 1 000 živě narozených dětí dle věku matky. Tyto hodnoty máme k dispozici od pěti různých skupin autorů a z různých let (celkem máme šest modelů, jak už jsme si řekli). Vzali jsme si tedy všechny tyto hodnoty a riziko jsme si vyjádřili na jedno narozené dítě pro každou věkovou kategorii matek. Jelikož naše hodnoty jsou pro věkové kategorie matek 15 – 44 let a věkovou kategorií 45⁺, musíme provést jednoduchou úpravu u hodnot rizika uvedených v referenčních studiích, neboť tam jsou věkové kategorie matek 15 – 55 let. Pro věkovou kategorií 45⁺ vezmeme průměr posledních hodnot rizika uvedených v referenčních studiích.

Abychom mohli provést porovnání, vybereme minimální a maximální hodnotu rizika pro jednotlivé věkové kategorie matek ze zmiňovaných šesti skupin hodnot rizika z referenčních studií. Situaci demonstrujeme na grafu 5.6, kam vynášíme minimální a maximální hodnoty rizika ze studií a hodnoty rizika z našeho modelu. Oranžová a červená linka představují minimální a maximální hodnoty a černá přerušovaná linka značí naše hodnoty. Riziko uvádíme v procentech. Můžeme si povšimnout, že naše hodnoty jsou nižší než je tomu ve světě, což je pro nás určitě uspokojivé zjištění. Ještě uvedeme druhý graf 5.7, který zobrazuje stejnou situaci s tím rozdílem, kdy místo hodnot rizika uvádíme logit rizika pro lepší čitelnost grafu.

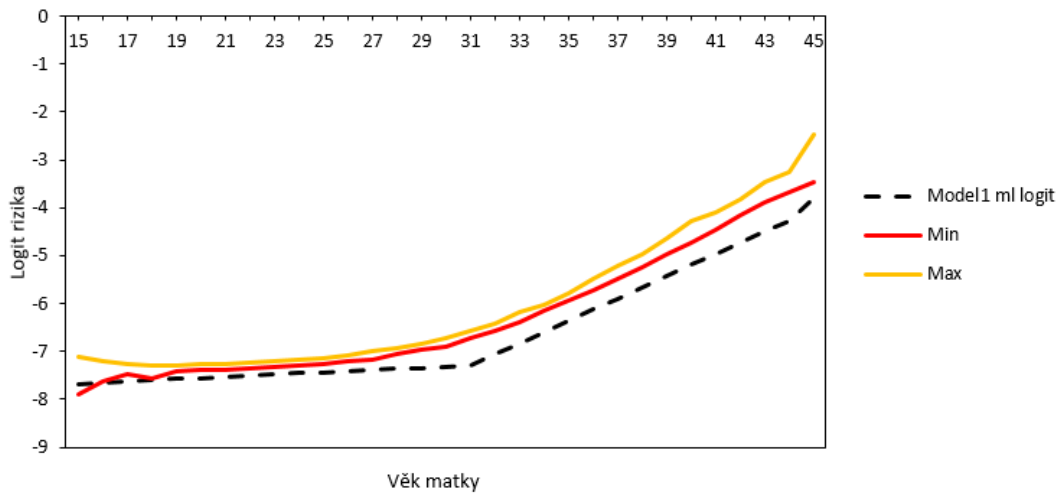
Table 1—Predicted prevalence of Down syndrome births per 1000 births

Age of mother at birth	Cuckle <i>et al.</i> , 1987	Hecht and Hook, 1994	Hecht and Hook, 1996	Bray <i>et al.</i> , 1998	Huether <i>et al.</i> , 1998	Morris <i>et al.</i> , 2002a
15	0.63	0.59	0.51	0.67	0.82	0.66
16	0.64	0.60	0.54	0.67	0.75	0.66
17	0.64	0.60	0.57	0.67	0.71	0.66
18	0.64	0.60	0.60	0.68	0.67	0.67
19	0.65	0.61	0.63	0.68	0.65	0.67
20	0.65	0.62	0.66	0.69	0.64	0.68
21	0.66	0.63	0.68	0.70	0.63	0.68
22	0.67	0.64	0.71	0.72	0.64	0.69
23	0.69	0.66	0.73	0.74	0.66	0.71
24	0.71	0.68	0.77	0.76	0.68	0.72
25	0.74	0.71	0.80	0.79	0.72	0.75
26	0.78	0.75	0.85	0.84	0.78	0.78
27	0.83	0.81	0.91	0.90	0.85	0.82
28	0.89	0.88	0.98	0.97	0.94	0.88
29	0.98	0.98	1.08	1.08	1.06	0.96
30	1.10	1.12	1.21	1.22	1.21	1.07
31	1.25	1.30	1.38	1.40	1.40	1.22
32	1.46	1.54	1.62	1.64	1.65	1.44
33	1.74	1.86	1.93	1.97	1.97	1.75
34	2.11	2.29	2.36	2.40	2.39	2.19
35	2.60	2.87	2.95	2.97	2.94	2.83
36	3.25	3.64	3.76	3.72	3.68	3.74
37	4.12	4.67	4.89	4.73	4.67	5.01
38	5.27	6.06	6.46	6.06	6.02	6.72
39	6.81	7.91	8.64	7.82	7.88	8.93
40	8.86	10.39	11.65	10.15	10.45	11.64
41	11.59	13.70	15.74	13.24	14.07	14.77
42	15.22	18.14	21.22	17.31	19.20	18.17
43	20.05	24.08	28.37	22.67	26.55	21.65
44	26.48	32.04	37.41	29.70	37.15	25.01
45	35.03	42.69	48.34	38.89	52.50	28.12
46	46.43	56.94	60.78	50.83	74.76	30.86
47	61.59	76.01	73.85	66.26	106.87	33.21
48	81.77	101.54	86.05	86.00	152.48	35.16
49	108.64	135.72	95.37	110.97	215.49	36.75
50	144.40	181.47	99.70	142.13	298.65	38.02
51	192.01	242.70	97.43	180.31	401.30	39.02
52	255.37	324.66	88.18	226.08	517.22	39.80
53	339.71	434.36	73.20	279.55	634.87	40.41
54	451.98	581.21	55.17	340.14	741.29	40.87
55	601.42	777.76	37.37	406.48	827.42	41.23

Obrázek 5.5: Tabulka - předpokládaná prevalence narozených Downů na 1 000 novorozenců (zahraniční referenční studie) [11]



Obrázek 5.6: Porovnání rizika podle věku matky našeho modelu a referenčních zahraničních studií (maximum a minimum)



Obrázek 5.7: Porovnání logitu rizika podle věku matky našeho modelu a referenčních zahraničních studií (maximum a minimum)

5.3.2. Řešení pomocí joinpoint regrese

Postup popsany výše je poměrně náročný v tom slova smyslu, že hledáme věk matky v , který bude co nejvýstižněji segmentovat křivku rizika. Znamená to tedy spoustu výpočtů a spoustu času. Oproti již zmíněnému postupu, pomocí logistické regrese, se může zdát joinpoint regrese daleko jednodušší. V programu R si nainstalujeme knihovnu `segmented`. Zadáme si počáteční odhad bodu zvratu ψ_0 a program de facto počítá za nás.

Nejprve si vyjádříme náš model ve tvaru 5.12

$$\text{logit}(p_v) = \begin{cases} a + bv & \text{pro } v \leq \psi, \\ a + bv + c(v - \psi) & \text{pro } v > \psi, \end{cases} \quad (5.22)$$

kde v je věk matky, a , b , c jsou neznámé regresní koeficienty a ψ je parametr. Model 5.22 můžeme přepsat do souhrnného tvaru

$$\text{logit}(p_v) = a + bv + c(v - \psi)I_{(v > \psi)}, \quad (5.23)$$

kde v je věk matky, a , b , c jsou neznámé regresní koeficienty, $I(\cdot)$ je identifikátor a ψ je parametr.

Pro pevně daný počáteční odhad ψ_0 bodu zvratu ψ můžeme model 5.23 odhadnout iterativním odhadováním lineárního modelu, který má tvar

$$a + bv + c(v - \psi)I_{(v > \psi)} \approx a + bv + c(v - \psi_0)I_{(v > \psi_0)} + \gamma I_{(v > \psi_0)}, \quad (5.24)$$

kde

$$\frac{\gamma}{c} = \frac{\psi_0 - \psi}{1},$$

tedy

$$\psi = \psi_0 - \frac{\gamma}{c}.$$

Přičemž v je věk matky, a , b , c jsou neznámé regresní koeficienty, $I(\cdot)$ je identifikátor a γ je parametr, který měří nespojitost v bodě zvratu ψ .

Po linearizaci odhadneme parametry modelu metodou maximální věrohodnosti. Udatujeme iteračně odhad parametru ψ , dokud nenajdeme optimální bod zvratu.

Výhodou joinpoint regrese oproti postupu, kdybychom hledali optimální odhad parametru ψ na „mřížce“, je to, že v joinpoint regresi můžeme určit i odhad nejistoty odhadu tohoto parametru.

Výpočet v programu R

Pro výpočet v programu R si označme [14], [19]

```
# age ... věk matky,  
# born ... celkový počet narozených dětí pro věkové kategorie matek,  
# born1 ... celkový počet narozených dětí pro věkové kategorie matek očištěný o podíl  
nezjištěných,  
# downsbyage ... počet narozených Downů pro věkové kategorie matek,  
# downsbyage1 ... podíl nezjištěných,  
# risk ... riziko výskytu DS pro věkové kategorie matek.
```

Tak jako v předešlém textu, budeme uvažovat věkové kategorie matek 15 – 44 let a věkovou kategorii matek 45⁺. Abychom mohli vypočítat born1, musíme nejdříve spočítat downsbyage1. downsbyage1 představuje podíl nezjištěných, tj. podíl Downů ve věkové kategorii matek, kdy věk matky nebyl zjištěný. Toto jsme řešili v kapitole 3, tak podrobný postup výpočtu nebudeme znovu uvádět. Tímto podílem poté upravíme celkový počet narozených dětí (počty se sníží).

```
> born1<-born*downsbyage1
```

Riziko pak vypočítáme jako

```
> risk<-downsbyage/born1
```

Teď už máme všechny pomocné výpočty k tomu, abychom v R mohli vypočítat odhad parametrů lineární lomené regresní funkce pomocí knihovny `segmented`. Nejprve

vycházíme z obvyklé regrese vypočtené příkazem `glm`. Parametr `family` specifikuje konkrétní rozdělení pravděpodobnosti modelované veličiny.

```
> model<-glm(risk~age, family=binomial(logit), weights=born1)
```

Nainstalujeme a následně načteme knihovnu `segmented`.

```
> install.packages("segmented")
```

```
> library("segmented")
```

Dále `seg.Z` označuje bod, ve kterém bude odhadován bod zlomu. Bod `psi` je počáteční odhad bodu zlomu. Tady je třeba podotknout, že tím je dán počet bodů zvratu. My tudíž budeme mít pouze jeden bod zvratu.

```
> model.s<-segmented(model, seg.Z=~age, psi=31)
```

Postupně vypíšeme a znázorníme vypočtené výsledky. Nejdříve máme

```
> slope(model.s)
```

Na obrázku 5.8 ve sloupci `Est` jsou odhadnuté směrnice jednotlivých úseků. Obsah sloupců je uveden v záhlaví – máme tam i vypočítané 95 % intervaly spolehlivosti. Podrobnější výpis získáme pomocí příkazu

```
> summary(model.s)
```

Na obrázku 5.8 vidíme získané výsledky. Odhadovaný bod zlomu `Estimated Break-Point(s)` nám vyšel ve věku matky 32,725, což je trochu jiný výsledek, než který jsme získali pomocí logistické regrese. Veličina `U1.age` popisuje, o kolik se musí změnit směrnice v bodě zlomu proti směrnici v předchozím úseku. Graficky je výsledek

znázorněn na obrázku 5.9. Využili jsme opět empirický logit rizika pro lepší čitelnost grafu. Příslušné příkazy jsou

```
> empir.logit=log(risk/(1-risk))
> plot(empir.logit~age, pch=20)
> plot.segmented(model.s, add=TRUE, col="red", lwd=2)
```

```
> slope(model.s)
$age
      Est.  St.Err. t value CI(95%).l CI(95%).u
slope1 0.03677 0.004552  8.077  0.02743  0.04611
slope2 0.26920 0.008745 30.790  0.25130  0.28720

> summary(model.s)

***Regression Model with Segmented Relationship(s)***

Call:
segmented.glm(obj = model, seg.Z = ~age, psi = 31)

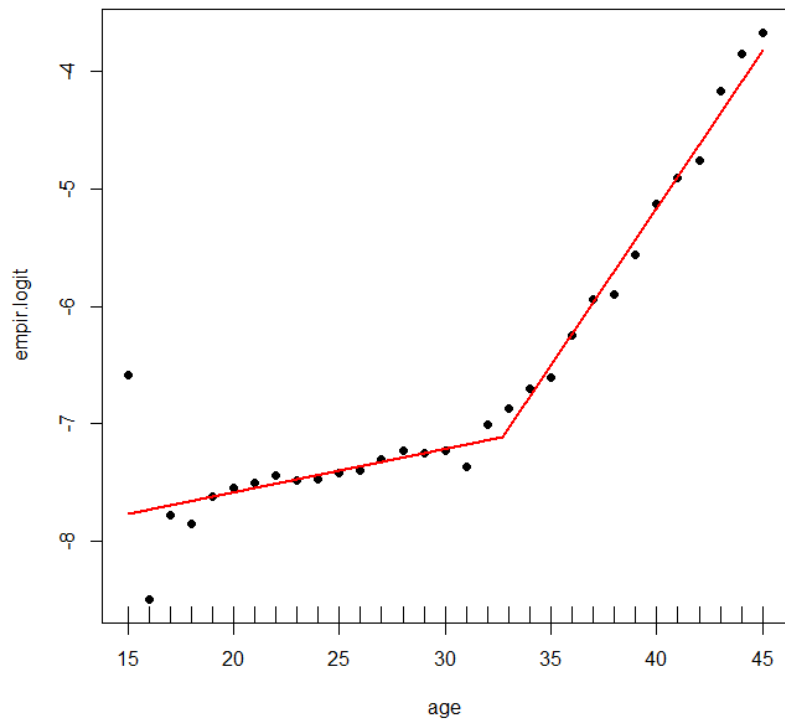
Estimated Break-Point(s):
      Est. St.Err
32.725  0.288

Meaningful coefficients of the linear terms:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.318906  0.113564 -73.253 < 2e-16 ***
age          0.036766  0.004552  8.077 6.66e-16 ***
U1.age       0.232460  0.009859 23.579      NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2169.833 on 30 degrees of freedom
Residual deviance: 41.119 on 27 degrees of freedom
AIC: 249.91

Convergence attained in 3 iterations with relative change 1.937075e-13
```

Obrázek 5.8: Výsledek funkce slope a summary pro joinpoint model s jedním bodem zvratu



Obrázek 5.9: Odhadnutý model vypočítaný v programu R. Znáznorňuje logit pravděpodobnosti výskytu Downova syndromu v závislosti na věku matky.

Kapitola 6

Standardizace

Na začátku diplomové práce jsme si kladli za cíl, že se pokusíme vysvětlit, zda nárůst počtu výskytů Downova syndromu je zapříčiněn vyšším věkem matek nebo je způsoben ještě nějakými dalšími faktory. V kapitole 5 jsme sestrojili modely závislosti rizika Downova syndromu na věku matky a mohli jsme si povšimnout, že závislost je po 31., respektive po 32., roce života matek daleko silnější než v nízkých věkových kategoriích (riziko roste rychleji). Nyní nás bude zajímat, zda ve sledovaném období let 1961 – 2010 došlo ke změně ve výskytu DS. Při hledání odpovědi, chceme být co nejméně závislí na tom, jaký model závislosti rizika Downova syndromu na věku matky použijeme. Další problém, který vyvstává, je, že do prenatalně diagnostikovaných Downů jsou započítáváni i ti, kteří by se možná nenarodili - došlo by ke spontánnímu potratu. Celkový počet Downů tak roste, i když zohledníme měnící se věkovou strukturu. K těmto účelům se používá věková standardizace, kdy se oprostíme od měnící se věkové struktury matek.

6.1. Obecný popis procesu standardizace

Standardizace umožňuje nezkrácené srovnání dvou či více populací lišících se svou strukturou. Smyslem je převést ukazatele porovnávaných populací na společný základ a tím odstranit nestejnorodost podmíněnou odlišnou strukturou. Společným základem pro přepočítávání je tzv. *standardní populace*. Na základě standardní populace se přepočítávají hodnoty hrubých ukazatelů (při výpočtu nepřihlíží ke struktuře zavádějících

faktorů) na společný základ. Standardizují se různé ukazatele jako je například úmrtnost, prevalence, porodnost, a to podle různých faktorů jako je například věk, pohlaví nebo stádium nemoci. Standardizace používá dvě základní metody – *přímá standardizace* a *nepřímá standardizace*. Konkrétní metodu standardizace vybíráme většinou podle toho, jaké údaje máme k dispozici. Nejčastěji se však používá metoda přímé standardizace.

Nyní si metodu přímé standardizace vysvětlíme na příkladu, kdy zavádějícím faktorem je věk a hrubým ukazatelem úmrtnost. Metodu nepřímé standardizace interpretovat nebudeme, jelikož jsme s ní nepracovali. Čerpáno z [4], [14], [24].

6.1.1. Přímá standardizace

Přímá standardizace na věkovou strukturu spočívá v tom, že se zvolí určitá populace, které se říká standardní, a hrubý ukazatel dané populace se přepočte vzhledem k její věkové struktuře. K tomu je nutné stanovit očekávaný počet úmrtí ve sledované populaci. To je hypotetický počet úmrtí, které bychom ve sledované populaci očekávali, kdyby se její věková struktura rovnala věkové struktuře standardní populace. Jinak řečeno, je to hypotetický počet úmrtí ve standardní populaci, u které platí věkově specifické úmrtnosti sledované populace.

Označíme-li c_i jako očekávaný počet úmrtí v i -té věkové kategorii pro $i = 1, \dots, k$, kde k označuje počet věkových kategorií, pak platí

$$c_i = p_i n_i^*, \quad (6.1)$$

kde p_i označuje specifickou úmrtnost v i -té věkové kategorii dané populace a n_i^* označuje četnost standardní populace v i -té věkové kategorii. Pro výpočet celkového počtu očekávaných úmrtí c platí vztah

$$c = \sum_{i=1}^k c_i = \sum_{i=1}^k p_i n_i^*. \quad (6.2)$$

Přímo standardizovaná úmrtnost SM_D na věkovou strukturu se vypočítá jako podíl celkového počtu očekávaných úmrtí c podělený celkovou četností n^* standardní popu-

lace

$$SM_D = \frac{c}{n^*}, \quad (6.3)$$

kde $n^* = \sum_{i=1}^k n_i^*$. Používání přímé standardizace není vhodné u malých populací, kde nemusí být dostatečně přesné specifické úmrtnosti p_i a tím pádem i odhad standardizovaného ukazatele.

Přímo standardizovaná úmrtnost, která se vypočítala z empirických dat, je bodovým odhadem (\widehat{SM}_D) populační přímo standardizované úmrtnosti. Nyní budeme chtít ještě určit intervalový odhad (neboli konfidenční interval). Předpokládejme, že počet úmrtí je veličina s binomickým rozdělením pravděpodobnosti. Pokud jsou počty úmrtí d_i ve věkových skupinách nezávislé, počty osob n_i a počty úmrtí d_i ve věkových skupinách jsou dostatečně velké, pak meze $100(1 - \alpha) \%$ intervalu spolehlivosti pro SM_D se vypočítají podle vzorce

$$(\widehat{SM}_D - z_{1-\frac{\alpha}{2}} SE(\widehat{SM}_D); \widehat{SM}_D + z_{1-\frac{\alpha}{2}} SE(\widehat{SM}_D)). \quad (6.4)$$

Směrodatnou odchylku $SE(\widehat{SM}_D)$ vypočítáme jako

$$SE(\widehat{SM}_D) = \frac{1}{n^*} \sqrt{\sum_{j=1}^k n_j^* \widehat{p}_j (1 - \widehat{p}_j) / n_j}, \quad (6.5)$$

kde

\widehat{p}_i ... odhady specifických úmrtností v dané populaci vypočítané jako

$$\widehat{p}_i = \frac{d_i}{n_i},$$

d_i ... počet úmrtí ve věkových skupinách,

n_i ... počet osob ve věkových skupinách.

Za $z_{1-\frac{\alpha}{2}}$ dosazujeme hodnotu $100(1 - \alpha/2) \%$ kvantilu standardizovaného normálního rozdělení.

6.2. Klouzavý průměr

V této podkapitole popíšeme metodu klouzavého průměru, která se používá k vyhlazení časové řady. Tuto metodu později využijeme na námi analyzovaná data. Než přistoupíme k popisu klouzavého průměru, připomeňme si, co je vlastně časová řada. Čerpáno z [1], [2], [6], [22].

Časová řada jsou věcně a prostorově srovnatelná pozorování, která jsou jednoznačně uspořádána z hlediska času ve směru minulost-přítomnost. Analýza časových řad je pak soubor metod, které slouží k popisu časové řady a předvídání budoucího chování časové řady. Jednotlivé metody mají svoji filosofii, předpoklady, výhody, nevýhody a dají se popřípadě kombinovat jednotlivé metody z různých přístupů. Dále budeme předpokládat, že čtenář má základní poznatky z teorie časových řad a metod jejich řešení.

Metoda klouzavých průměrů se řadí mezi adaptivní metody pro konstrukci trendu a vyhlazování časové řady. Adaptivní metoda dokáže pracovat s trendem, který v čase mění svůj charakter a který tedy nelze popsat jedinou křivkou. Na druhé straně se však předpokládá, že v krátkých úsecích časové řady je možné vyrovnání pomocí matematických křivek i přesto, že tyto křivky mají většinou v různých úsecích odlišné parametry. Je tudíž možné, pouze lokálně vyrovnat trend. Výpočty probíhají rekurentně pomocí adaptačních vzorců a to vždy, když přibude nové pozorování.

Základní myšlenkou metody klouzavých průměrů je, že nejdříve zvolíme délku okna m . Zvolené okno postupně „klouže“ po časové řadě. Poté data v okně proložíme polynomem, tzv. *trendovou funkcí* (prokládání dat polynomem je to stejné jako vážené průměry v rámci okna) a spočítáme vyrovnanou hodnotu pro střed okna. Pak posuneme okno doprava a opět se vrátíme ke kroku, kdy data prokládáme polynomem.

Budeme předpokládat, že pracujeme s časovou řadou ve tvaru

$$y_t = T_t + \varepsilon_t, \quad t = 1, \dots, n.$$

Vyhlazování časové řady znamená, že tedy budeme odhadovat trend $\hat{y}_t = \hat{T}_t$. Při konstantním a lineárním trendu mluvíme o prostém klouzavém průměru, pokud je trend

kvadratický nebo polynomiální vyššího řádu, tak mluvíme o váženém klouzavém průměru.

Rozlišujeme dva typy klouzavého průměru

- *necentrováný* (obyčejný) klouzavý průměr – délka okna m je lichá,
- *centrováný* klouzavý průměr – délka okna m je sudá, přidáváme navíc krok centrování. Zprůměrováním dvou sousedních vyrovnaných hodnot se dostaneme k původním celočíselným časovým okamžikům.

Zaved' me si následující značení, kdy místo indexu t , budeme psát (t, i)

t ... střed okna,

i ... index pozice vůči středu.

$$t = \begin{cases} p+1, p+2, \dots, n-p & \text{pro } m = 2p+1 \\ p+\frac{1}{2}, p+\frac{3}{2}, \dots, n-p+\frac{1}{2} & \text{pro } m = 2p \end{cases}$$

$$i = \begin{cases} -p, \dots, -1, 0, 1, \dots, p & \text{pro } m = 2p+1 \\ -p+\frac{1}{2}, \dots, -\frac{1}{2}, \frac{1}{2}, \dots, p-\frac{1}{2} & \text{pro } m = 2p \end{cases}$$

Střed okna je roven $i = 0$ a platí

$$y_t = y_{t+i}$$

U vážených klouzavých průměrů se odhad trendu vypočítá jako

$$\hat{y}_t = \sum_i w_i y_{t,i}$$

kde w_i jsou příslušné váhy. Jednoduché klouzavé průměry, kdy trendy na klouzavých úsecích $2p+1$ hodnot jsou konstantní nebo lineární, mají všechna pozorování v okně stejnou váhu, tj. $w_i = \frac{1}{m}$ pro $\forall i$.

Uvažujme polynomiální trend ve tvaru

$$T_{t,i} = a_{0t} + a_{1t}i + a_{2t}i^2 + \dots + a_{rt}i^r,$$

r ... řád klouzavého průměru,
 a_{0t}, \dots, a_{rt} ... parametry modelu,
 i ... centrováný čas v okně,
 $T_{t,0} = a_{0t}$.

Odhady parametrů polynomického trendu získáme metodou nejmenších čtverců minimalizací součtu

$$\sum_i (y_{t,i} - a_{0t} - a_{1t}i - a_{2t}i^2 - \dots - a_{rt}i^r)^2 \rightarrow \min.$$

Řešením jsou normální rovnice

$$\left. \begin{aligned} \frac{\partial}{\partial a_0} : -2 \sum_i (y_{t,i} - a_{0t} - a_{1t}i - a_{2t}i^2 - \dots - a_{rt}i^r) &\stackrel{!}{=} 0 \\ \frac{\partial}{\partial a_1} : -2 \sum_i (y_{t,i} - a_{0t} - a_{1t}i - a_{2t}i^2 - \dots - a_{rt}i^r) &\stackrel{!}{=} 0 \\ \dots \dots \dots \\ \frac{\partial}{\partial a_r} : -2 \sum_i (y_{t,i} - a_{0t} - a_{1t}i - a_{2t}i^2 - \dots - a_{rt}i^r) &\stackrel{!}{=} 0 \end{aligned} \right\} r + 1 \text{ normálních rovnic}$$

$$\sum_i y_{t,i} = a_{0,t} \sum 1 + a_{1t} \sum i + a_{2t} \sum i^2 + \dots + a_{rt} \sum i^r \quad (0)$$

$$\sum_i i y_{t,i} = a_{0,t} \sum i + a_{1t} \sum i^2 + a_{2t} \sum i^3 + \dots + a_{rt} \sum i^{r+1} \quad (1)$$

.....

$$\sum_i i^r y_{t,i} = a_{0,t} \sum i^r + a_{1t} \sum i^{r+1} + a_{2t} \sum i^{r+2} + \dots + a_{rt} \sum i^{2r} \quad (r)$$

Pokud máme centrováný čas, tak platí

$$\sum_i i^k = 0 \text{ pro } k = 1, 3, 5, \dots$$

Pro výpočet parametrů $a_{0t}, a_{2t}, a_{4t}, \dots$ využijeme sudé rovnice (0), (2), (4),
 Odhady jsou lineární kombinace výrazů $\sum_i i^k y_{t,i}$ pro k sudé. Platí $i^k = (-i)^k$

pro $k = 0, 2, 4, \dots$. Odhady jsou pak ve tvaru $\hat{a}_{lt} = \sum_i c_i^{(l)} y_{t,i}$ pro $l = 0, 2, 4, \dots$ a $c_i^{(l)}$ jsou symetrické váhy vůči i . Pak odhad $\hat{a}_{0t} = \sum_i c_i^{(0)}$, kde $c_i^{(0)} = w_i$.

$c_i^{(0)} = w_i$ jsou váhy symetrické, normované, nezávislé na t , závislé na r a m .

Pokud je r sudé, pak na výpočet a_{0t} nemá vliv následující $r + 1$ -ní rovnice. To znamená, že odhad \hat{a}_{0t} je stejný pro řád r i $r + 1$.

$\hat{y}_t = \hat{y}_{t,0} = \hat{a}_{0t}$ jsou vyrovnané hodnoty ve středu okna, stačí odhadnout absolutní člen.

Pro výpočet parametrů $a_{1t}, a_{3t}, a_{5t}, \dots$ využijeme liché rovnice (1), (3), (5), \dots

Pak odhady parametrů jsou lineární kombinace výrazů $\sum_i i^k y_{t,i}$ pro k liché. Platí

$i^k = -(-i^k)$ pro $k = 1, 3, 5, \dots$

Odhady jsou pak ve tvaru $\hat{a}_{lt} = \sum_i c_i^{(l)} y_{t,i}$ pro $l = 1, 3, 5, \dots$

Výsledkem použití metody klouzavých průměrů je odhad trendové složky časové řady. Avšak její nevýhodou je, že prvních a posledních m hodnot trendové složky není určeno. Uvedeme si tedy způsob výpočtu těchto počátečních a koncových hodnot, které využívají jiné sady vah a jiný postup. Pro vyrovnané hodnoty platí

$$\hat{y}_{t,j} = \sum_{l=0}^r \hat{a}_{lt} j^l = \sum_{l=0}^r \sum_i c_i^{(l)} y_{t,i} j^l = \sum_i \sum_{l=0}^r c_i^{(l)} j^l y_{t,i} = \sum_i v_i^{(j)} y_{t,i},$$

$v_i^{(j)}$... váhy pro vyrovnané hodnoty s pozicí j , platí pro ně $v_i^{(j)} = v_{-i}^{(-j)}$, nejsou symetrické ani normované, závisí na r, m, j .

Počáteční hodnoty

1.okno

$$t = \begin{cases} p + 1 & \text{pro } m \text{ liché, } m = 2p + 1 \\ p + \frac{1}{2} & \text{pro } m \text{ sudé, } m = 2p \end{cases}$$

levá část okna

$$j = \begin{cases} -p, -p + 1, \dots, -1 & \text{pro } m \text{ liché} \\ -p + \frac{1}{2}, \dots, \frac{1}{2} & \text{pro } m \text{ sudé} \end{cases}$$

Koncové hodnoty

poslední okno

$$t = \begin{cases} n - p & \text{pro } m \text{ liché, } m = 2p + 1 \\ n - p + \frac{1}{2} & \text{pro } m \text{ sudé, } m = 2p \end{cases}$$

pravá část okna

$$j = \begin{cases} 1, \dots, p & \text{pro } m \text{ liché} \\ \frac{1}{2}, \frac{3}{2}, \dots, p - \frac{1}{2} & \text{pro } m \text{ sudé} \end{cases}$$

Pokud chceme předpověď časové řady, tak prodloužíme trend z posledního okna

$$t = \begin{cases} n - p & \text{pro } m \text{ liché, } m = 2p + 1 \\ n - p + \frac{1}{2} & \text{pro } m \text{ sudé, } m = 2p \end{cases}$$

budoucnost

$$j = \begin{cases} p + 1, p + 2, \dots & \text{pro } m \text{ liché} \\ p + \frac{1}{2}, p + \frac{3}{2}, \dots & \text{pro } m \text{ sudé} \end{cases}$$

Poslední otázkou, co zůstává pro metodu klouzavých průměrů, je, jaký řád a délku klouzavého průměru zvolit pro danou časovou řadu. Obvykle musíme učinit subjektivní posouzení charakteru dat, což nemusí být vždy jednoduché. Preferujeme jednoduché průměry co nejnižšího řádu a délku okna m volíme podle požadovaného stupně vyhlazení řady. Platí pravidlo, že čím větší je délka okna, tím větší je vyhlazení časové řady.

6.3. Aplikace na námi analyzovaná data

Když už jsme si popsali, v čem spočívá princip standardizace, můžeme přistoupit k aplikaci na námi analyzovaná data.

Jak už bylo vyřčeno, standardizace umožňuje nezkreslené srovnání dvou populací lišících se svou věkovou strukturou. S rostoucím věkem se riziko výskytu Downova syndromu zvyšuje, ukazatel typického počtu Downů je výrazně závislý na věku

matky. Někdy se k porovnávání dvou populací využívá hrubá míra, což je nejjednodušší ukazatel. Hrubá míra vyjadřuje úroveň počtu DS. Je definována jako podíl narozených Downů na 100 000 narozených dětí. Nevýhodou ukazatele je skutečnost, že nezohledňuje věková specifika dané populace. Ukazuje tedy pouze celkové počty DS bez ohledu na věkovou strukturu populace matek. Proto se používá věkově standardizovaná míra, což je vážený průměr hrubých měr vypočtených pro různé věkové skupiny matek a váhami jsou podíly jednotlivých věkových skupin ve standardní populaci matek. Čerpáno z [4] a [24].

Budeme pracovat s metodou přímé standardizace, kdy jsou věkově specifické počty DS reálné populace aplikovány na populaci standardní. V našem případě si za standardní populaci zvolíme populaci s věkovou strukturou, jaká byla v roce 2010. Standardizované počty DS se běžně uvádí v přepočtu na 100 000 narozených dětí. Přepočtem podle příslušného vztahu získáme hodnotu počtu DS, která by se vyskytovala v reálné populaci za předpokladu, že její věková struktura by odpovídala věkové struktuře populace standardní. Čili zajímá nás, kolik Downů by bylo například v roce 1961, kdyby tehdy byla věková struktura jako v roce 2010.

Výše jsme si napsali něco k základnímu principu standardizace a co nás bude zajímat. Nyní přistoupíme k samotnému postupu jejího výpočtu pro námi analyzovaná data. Dělali jsme standardizaci pro dva případy. Nejprve pro celkový počet narozených Downů a jako druhý případ pro součet narozených Downů a počet prenatalně diagnostikovaných, kteří se nenarodili (tedy počet přerušovaných těhotenství postižených Downovým syndromem). Začneme standardizací pro celkový počet narozených Downů.

Nejprve zavedme následující značení. Pracujeme s údaji za 1 rok, a proto roky neindexujeme.

D_i ... počet Downů v daném roce v i -té věkové kategorii matek,

N_i ... počet narozených dětí v daném roce v i -té věkové kategorii matek,

U ... počet narozených Downů v daném roce u matek s nezjištěným věkem,

C ... celkový počet všech narozených Downů,

i ... index věku matky.

Celkový počet všech narozených Downů se počítá jako

$$C = \sum_i D_i + U.$$

Hrubou míru pro počet Downů v i -té věkové kategorii matek (počet Downů na 100 000 narozených dětí v daném roce) vypočítáme podle vztahu

$$R_i = \frac{D_i}{\left(1 - \frac{U}{C}\right)} \frac{100\,000}{N_i}. \quad (6.6)$$

Dále označme

ST_i ... velikost populace matek i -té věkové kategorie v referenčním roce 2010.

Pak věkově standardizovaná míra (počet Downů na 100 000 narozených dětí v struktuře roku 2010) se rovná

$$AR = \sum_i \left(R_i \frac{ST_i}{\sum_j ST_j} \right). \quad (6.7)$$

Dále vypočítáme interval spolehlivosti věkově standardizované míry pro bodový odhad \widehat{AR} . Horní a dolní hranice intervalu spolehlivosti jsou

$$CI_{high} = \widehat{AR} + z_{1-\frac{\alpha}{2}} SE(\widehat{AR}),$$

$$CI_{low} = \widehat{AR} - z_{1-\frac{\alpha}{2}} SE(\widehat{AR}),$$

CI_{high} ... horní hranice intervalu spolehlivosti,

CI_{low} ... dolní hranice intervalu spolehlivosti,

$z_{1-\frac{\alpha}{2}}$... kvantil normovaného normálního rozložení pravděpodobnosti při hladině spolehlivosti $\alpha = 0,05$,¹

¹Vycházíme z aproximace normálním rozdělením na základě Moivre-Laplaceovy centrální limitní věty, protože počet událostí je vysoký.

$SE(\widehat{AR})$... směrodatná odchylka (střední chyba odhadu věkově standardizované míry).

Pro odvození směrodatné odchylky budeme nejdříve uvažovat obecný případ a využijeme informace z obecného popisu procesu přímé standardizace. Nejprve označme pro daný rok

X_i ... počet diagnóz DS u dětí matek v i -té věkové kategorii matek ve standardní populaci.

Předpokládejme, že počet diagnóz DS je náhodná veličina s binomickým rozdělením pravděpodobnosti $X_i \sim B_i(s_i, p_i)$, X_i vzájemně nezávislé náhodné veličiny. [13]

s_i ... počet narozených dětí matek v i -té věkové kategorii matek ve standardní populaci,
 p_i ... pravděpodobnost výskytu Downova syndromu v i -té věkové kategorii matek.

Sledujeme $\sum_i X_i$. Jelikož máme náhodnou veličinu s binomickým rozdělením pravděpodobnosti, tak rozptyl je ve tvaru

$$\text{var}\left(\sum_i X_i\right) = \sum_i \text{var}(X_i) = \sum_i s_i p_i (1 - p_i).$$

Dále určíme odhad rozptylu, neboť chceme určit směrodatnou odchylku bodového odhadu

$$\widehat{\text{var}}\left(\sum_i X_i\right) = \sum_i s_i \widehat{p}_i (1 - \widehat{p}_i),$$

kde $\widehat{p}_i = \frac{d_i}{n_i}$,

d_i ... skutečný počet diagnóz DS,

n_i ... skutečný počet narozených dětí v dané věkové kategorii matek v daném roce.

Nyní chceme určit směrodatnou odchylku věkově standardizované míry pro bodový odhad \widehat{AR} .

Rozptyl pro odhad \widehat{AR} určíme

$$\text{var}(\widehat{AR}) = \text{var} \left[\sum_i \left(\widehat{R}_i \frac{ST_i}{\sum_j ST_j} \right) \right] = \sum_i \left[\text{var} \left(\widehat{R}_i \frac{ST_i}{\sum_j ST_j} \right) \right] = \sum_i \left(\frac{ST_i}{\sum_j ST_j} \right)^2 \text{var}(\widehat{R}_i). \quad (6.8)$$

Označme

$$\widehat{R}_i = \frac{D_i}{(1 - \frac{U}{C})} = \frac{Y_i}{N_i}. \quad (6.9)$$

Předpokládejme

$$1 - \widehat{R}_i = 1 - \frac{Y_i}{N_i} = \frac{N_i - Y_i}{N_i} \approx 1. \quad (6.10)$$

Rozptyl $\text{var}(\widehat{R}_i)$ odhadneme jako

$$\widehat{\text{var}}(\widehat{R}_i) = \frac{1}{N_i^2} N_i \widehat{R}_i (1 - \widehat{R}_i) \approx \frac{1}{N_i^2} N_i \frac{Y_i}{N_i} 1 = \frac{Y_i}{N_i^2}. \quad (6.11)$$

Dosazením do vztahu 6.8 dostáváme odhad rozptylu ve tvaru

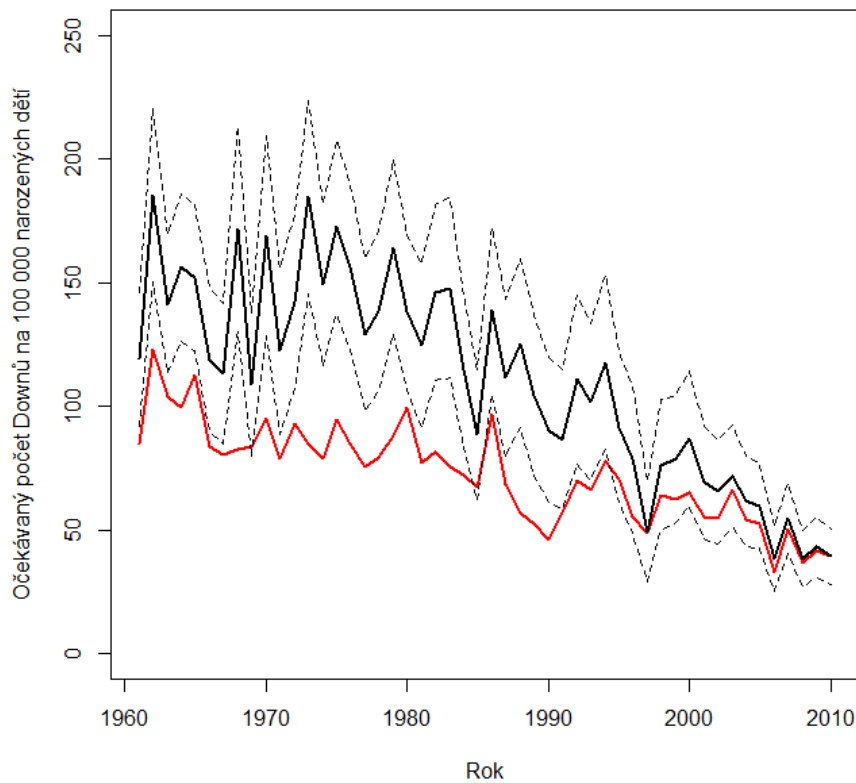
$$\widehat{\text{var}}(\widehat{AR}) = \sum_i \left(\frac{ST_i}{\sum_j ST_j} \right)^2 \frac{D_i}{(1 - \frac{U}{C})} \frac{1}{N_i^2}. \quad (6.12)$$

Směrodatnou odchylku $SE(\widehat{AR})$ vypočítáme jako druhou odmocninu rozptylu násobenou 100 000 (provádíme přepočít na 100 000 narozených dětí)

$$SE(100\,000 \widehat{AR}) = 100\,000 \left[\sum_i \left(\frac{ST_i}{\sum_j ST_j} \right)^2 \left(\frac{D_i}{(1 - \frac{U}{C})} \frac{1}{N_i^2} \right) \right]^{1/2}. \quad (6.13)$$

Výsledek standardizace vidíme na obrázku 6.1. Červená linka představuje počet Downů na 100 000 narozených dětí v daném roce, černá linka reprezentuje počet Downů na 100 000 narozených dětí při věkové struktuře roku 2010 a černé přerušované linky jsou intervaly spolehlivosti (konfidenční intervaly). Vidíme, že očekávané počty Downů jsou vyšší, když zohledníme měnící se věkovou strukturu matek. Průběh grafu je velice bouřlivý, těžko tedy můžeme konstatovat nějaké podrobnější informace o vývoji

očekávaného počtu DS.



Obrázek 6.1: Vývoj standardizovaných počtů narozených DS (1961-2010)

Standardizace pro součet narozených Downů a počet prenatalně diagnostikovaných, kteří se nenarodili (počet přerušných těhotenství postižených DS) se počítá obdobně. K výše uvedenému značení ještě dále označme

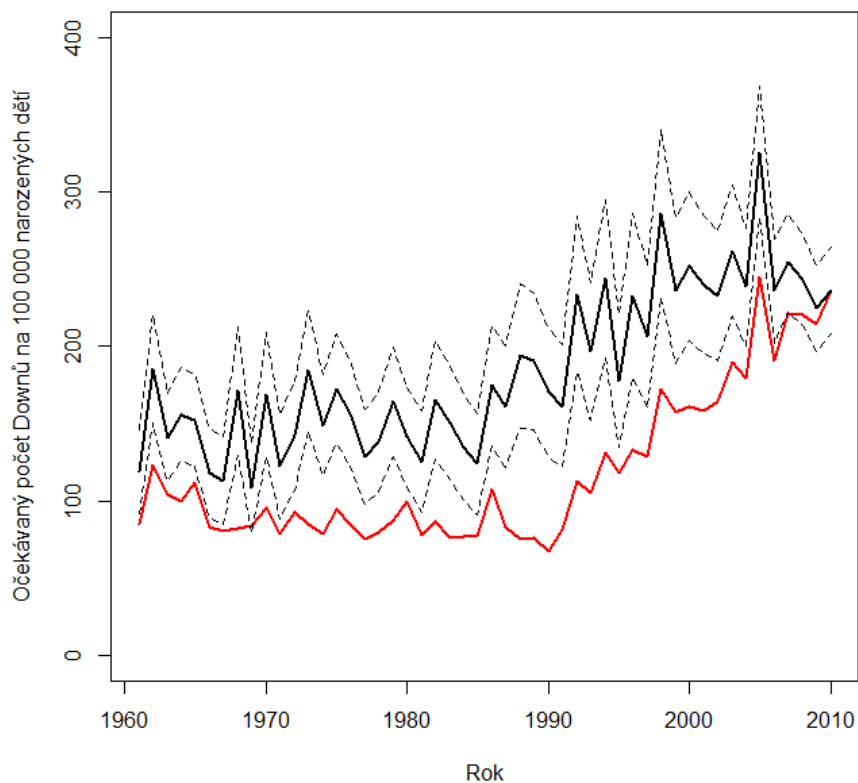
P_i ... počet prenatalně diagnostikovaných Downů v daném roce v i -té věkové kategorii matek,

UP ... počet prenatalně diagnostikovaných Downů v daném roce u matek s nezjištěným věkem.

Hrubou mírou pro součet narozených Downů a počet prenatalně diagnostikovaných, kteří se nenarodili, v i -té věkové kategorii matek vypočítáme podle vztahu

$$R_i = \frac{\frac{D_i + P_i}{\left(1 - \frac{U + UP}{C + \sum_i P_i}\right)}}{N_i} 100\,000. \quad (6.14)$$

Dále je postup analogický s výše uvedeným postupem, proto jej uvádět nebudeme.



Obrázek 6.2: Vývoj standardizovaných počtů diagnostikovaných (prenatálně i postnatálně) DS (1961-2010)

Výsledek druhé standardizace je na obrázku 6.2. Opět červená linka představuje počet Downů na 100000 narozených dětí v daném roce, černá linka vyjadřuje počet

Downů na 100 000 narozených dětí při věkové struktuře roku 2010 a černé přerušované linky jsou intervaly spolehlivosti. Nyní jsme do standardizace zohlednili počty narozených Downů a počty prenatalně diagnostikovaných Downů, kteří by se nenarodili. I zde můžeme pozorovat, že očekávané počty Downů jsou vyšší, pokud zohledníme měnící se věkovou strukturu matek. I na tomto grafu se nedají vyčíst nějaké bližší informace kvůli jeho velice neklidnému průběhu.

Vyhlazení klouzavým průměrem

V předchozí části této kapitoly jsme vyjádřili standardizované počty pro počet narozených Downů a také standardizované počty pro součet narozených Downů a počet prenatalně diagnostikovaných, kteří se nenarodili. Výsledky obou standardizací nám ukazují obrázky 6.1 a 6.2. Průběh křivek na obou obrázcích je velmi bouřlivý a nestabilní, je tedy těžké vyvozovat nějaké závěry, které jsme získali pomocí standardizace. Abychom tak mohli učinit, je třeba křivky vyhladit klouzavým průměrem.

Podíváme-li se na obrázek 6.1, vidíme průběh křivky za období 1961 – 2010. Tuto křivku tedy bereme jako časovou řadu, kterou budeme chtít vyhladit. K tomuto účelu použijeme metodu klouzavého průměru popsanou v podkapitole 6.2. Použijeme jednoduchý klouzavý průměr s délkou okna $m = 3$. Délka okna je lichá, nebude tudíž třeba centrování. Začneme případem celkového počtu narozených Downů. Výše uvedené vztahy postupně upravíme, budeme uvažovat stejné značení. Klouzavý průměr používáme jednak pro samotnou časovou řadu (suma od $j - 1$ do $j + 1$), ale také děláme průměr přes sousední věkové kategorie matek (suma od $i - 1$ do $i + 1$).

Celkový počet všech narozených Downů se počítá jako

$$C_j = \sum_i D_{ij} + U_j.$$

Hrubou míru pro počet Downů v populaci dětí (počet Downů na 100 000 naro-

zených dětí v daném roce) vypočítáme podle vztahu

$$R_{ij} = \frac{\sum_{i-1}^{i+1} \sum_{j-1}^{j+1} \frac{D_{ij}}{(1 - \frac{U_j}{C_j})}}{\sum_{i-1}^{i+1} \sum_{j-1}^{j+1} N_{ij}} 100\,000 \quad \text{pro } i = 2, \dots, v-1; j = 2, \dots, 49, \quad (6.15)$$

kde nyní

R_{ij} ... hrubá míra v j -tém roce v i -té věkové kategorii matek,

D_{ij} ... počet Downů v j -tém roce v i -té věkové kategorii matek,

N_{ij} ... počet matek v j -tém roce v i -té věkové kategorii matek,

U_j ... počet narozených Downů v j -tém roce u matek s nezjištěným věkem,

C_j ... celkový počet všech narozených Downů v j -tém roce,

i ... index věku matky,

j ... index roku,

v ... počet věkových kategorií matek, které uvažujeme.

Ze vztahu 6.15 vidíme, že počítáme bez prvního a posledního roku a bez první a poslední věkové kategorie.

Pak věkově standardizovaná míra (pro počet Downů na 100000 v struktuře roku 2010) se počítá jako ²

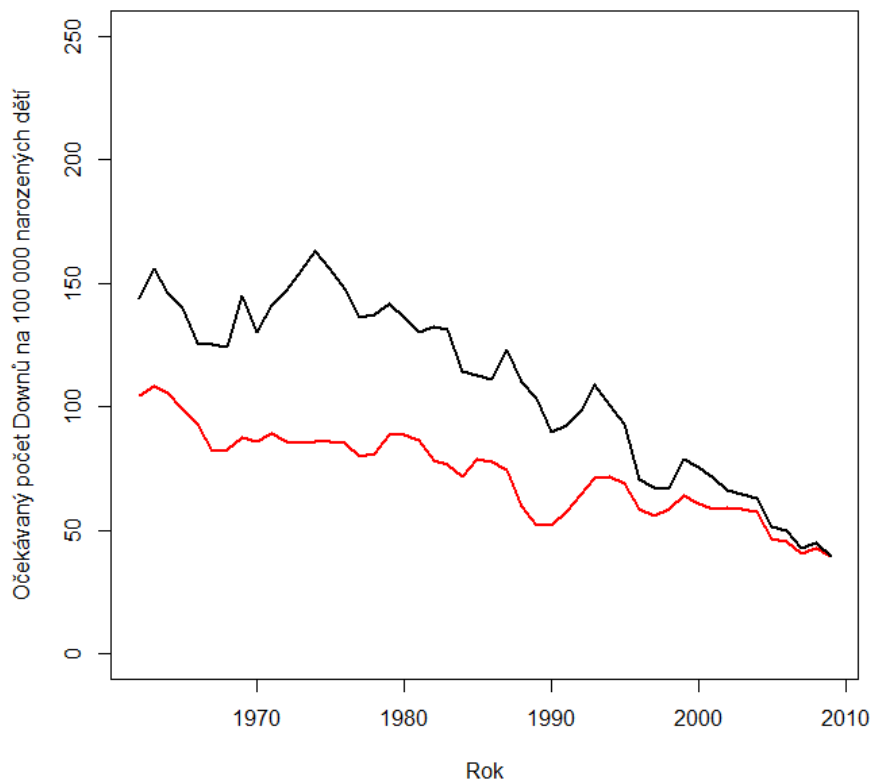
$$AR_j = \sum_{i=2}^{v-1} \left(R_{ij} \frac{ST_i}{\sum_{k=2}^{v-1} ST_k} \right). \quad (6.16)$$

Interval spolehlivosti nebudeme uvádět kvůli poměrně náročnému výpočtu směrodatné odchylky. V tomto případě směrodatnou odchylku nemůžeme počítat jako v 6.8, neboť máme chybějící nezávislost odhadu jednotlivých rizik, která pak váženě průměrujeme. Děláme-li tedy rozptyl součtu, nemůžeme použít jen součet rozptylu, ale je třeba tam přidat i kovarianční členy.

Výsledek vidíme na obrázku 6.3. Červená linka zobrazuje počet Downů na 100000 narozených dětí v daném roce, černá linka představuje počet Downů na 100000 narozených dětí při věkové struktuře roku 2010. Z obrázku je patrné (sledujeme černou

²Krajní hodnoty R_{ij} jsme nevyhlazovali.

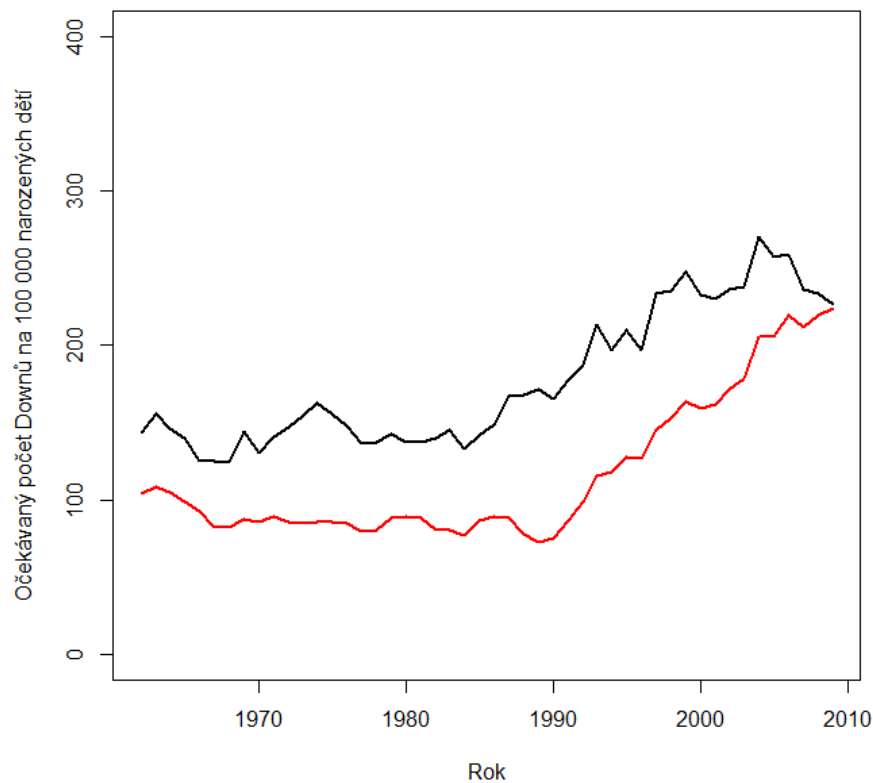
křivku), že od počátku 60. let do konce 80. let je počet narozených dětí s DS poměrně ustálený. Tento počet Downů odpovídá počtu skutečně narozených dětí. V průběhu 90. let 20. století docházelo k poklesu počtu postnatálních diagnóz (tedy skutečně narozených) dětí s DS, protože rostl význam prenatální diagnostiky a bylo zachyceno více případů těhotenství postižených DS (prenatálně diagnostikované, kteří se nenarodili, zachycuje druhá situace standardizace).



Obrázek 6.3: Vývoj standardizovaných počtů narozených DS (1961-2010) - vyhlazení klouzavým průměrem

V druhém případě, kdy jsme měli součet narozených Downů a počet prenatálně diagnostikovaných, kteří se nenarodili, se opět postupuje zcela analogicky, proto postup uvádět nebudeme. Ukážeme si až výsledek, který můžeme vidět na obrázku 6.4. Zase

červená linka reprezentuje počet Downů na 100 000 narozených dětí v daném roce, černá linka vyjadřuje počet Downů na 100 000 narozených dětí při věkové struktuře roku 2010. Z obrázku můžeme vyčíst (sledujeme černou křivku), že počet diagnóz Downova syndromu od počátku 60. let do konce 80. let je celkem stabilní. Tento počet DS odpovídá počtu narozených dětí. Dále vidíme, že v průběhu 90. let 20. století docházelo k růstu počtu diagnóz Downova syndromu. Tyto počty jsou větší než by vyjadřoval prostý zvyšující se věk matek. V posledním desetiletí již téměř nedocházelo k dalšímu přibývání diagnóz Downova syndromu.



Obrázek 6.4: Vývoj standardizovaných počtů diagnostikovaných (prenatálně i postnatálně) DS (1961-2010) - vyhlazení klouzavým průměrem

Závěr

V dnešním světě se potýkáme se spoustou nemocí či všelijakých vrozených vad. Jednou takovouto vrozenou vadou je Downův syndrom, který se nedá vyléčit. Je všeobecně známo, že riziko výskytu Downova syndromu roste s věkem matky. Věk matky je tedy zásadní faktor pro výskyt tohoto postižení. Dalším všeobecně známým faktem je, že pokrok jde neustále kupředu a to platí i v medicíně. Tudíž prenatální diagnostika používá nové metody, je včasnější a dokáže zachytit více těhotenství, která jsou postižená Downovým syndromem. Pak záleží na matce či rodičích dítěte, zda si postižené miminko ponechají a budou mu věnovat potřebnou péči. Prokáže-li se u nenarozeného dítěte toto postižení, má matka ze zákona nárok na umělé ukončení těhotenství.

Právě problematikou Downova syndromu a jeho závislosti na věku matky se zabírala tato diplomová práce. Hlavní náplní práce bylo vytvořit model, který by vyjádřil závislost rizika výskytu Downova syndromu na věku matky. Pro tento účel jsme měli data pro českou populaci matek za období 1961 – 2010. Byl to úkol složitý, neboť na počátku práce jsme neměli tušení, jakou metodu použít. Ve třetí kapitole jsme vykreslili spoustu grafů a vyčetli jsme mnoho podstatných skutečností. Pozorovali jsme, že opravdu prenatální diagnostika má od 90. let 20. století neopomenutelný význam a díky ní bylo zachyceno mnoho postižených těhotenství a tudíž počet narozených dětí trpících Downovým syndromem klesal. Od 80. let 20. století zároveň však klesá i celkový počet narozených dětí v České republice. Avšak od 90. let 20. století se rapidně zvýšil průměrný věk matek. Sledovali jsme i vývoj incidence Downova syndromu a mohli jsme pozorovat, že zhruba do roku 1990 byl vývoj víceméně stabilní, ale poté výrazně rostl. Vše tedy naznačovalo tomu, že tam určitá souvislost s věkem matky bude.

Stěžejním cílem diplomové práce bylo vytvoření modelu závislosti rizika výskytu Downova syndromu na věku matky. K tomuto účelu jsme použili dvě metody a to metodu logistické regrese a metodu joinpoint regrese. Nejdříve bylo tedy třeba teoreticky nastudovat tuto problematiku a následně ji uplatnit na námi analyzovaná data. Metodou logistické regrese jsme se dopracovali k tomu, že v našem modelu bude zlomovým bodem věk matky 31 let. Výsledek můžeme vidět na obrázku 5.4. Metoda joinpoint regrese byla pro nás zcela nová. Po nastudování patřičné teorie jsme za pomoci programu R vytvořili druhý model, který je znázorněn na obrázku 5.9. Pomocí této metody však dostáváme nepatrně jiný výsledek. Zlomovým bodem je nyní věk matky 32 let. I přes tyto dva trochu odlišné výsledky můžeme konstatovat, že ke změně závislosti rizika na věku matky došlo. Tato závislost je po 31., respektive 32., roce života matek daleko silnější než v nízkých věkových kategoriích (riziko roste rychleji).

Otázka, kterou jsme si dále kladli, byla, zda ve sledovaném období let 1961 – 2010 došlo ke změně ve výskytu DS. Při hledání odpovědi na tuto otázku jsme chtěli být co nejméně závislí na tom, který model závislosti rizika výskytu Downova syndromu na věku matky použít. K tomu jsme využili metodu standardizace. Díky standardizaci jsme se oprostili od závislosti na věku matky a zjistili jsme, že nejenom věk matky je důležitý faktor, ale že dochází i ke změně ve výskytu DS ve sledovaném období let 1961 – 2010. Na tuto skutečnost poukazuje obrázek 6.4, kde jsme standardizované počty nakonec ještě vyhladili klouzavým průměrem pro lepší ilustraci.

Literatura

- [1] ARLT, J., ARLTOVÁ, M., RUBLÍKOVÁ, E. *Analýza ekonomických časových řad s příklady*. 1. vyd., Praha: Vysoká škola ekonomická v Praze, 2002, ISBN 80-245-0307-7.
- [2] ARLT, J. *Moderní metody modelování ekonomických časových řad*. 1. vyd., Praha: Grada, 1999, ISBN 80-7169-539-4.
- [3] BAZALOVÁ, B., TOMIŠKOVÁ, L. *Život, výchova a vývoj jedince s Downovým syndromem v rodině*. 2006.
- [4] BENCKO, V. a kol. *Statistické metody v epidemiologii*. 1. vyd., Praha: Karolinum, 2003, 236 s., Biomedicínská statistika, ISBN 80-246-0763-8.
- [5] BUDÍKOVÁ, M., KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. 1. vyd., Praha: Grada, 2010, 272 s., Expert (Grada), ISBN 978-80-247-3243-5.
- [6] CIPRA, T. *Analýza časových řad s aplikacemi v ekonomii*. 1. vyd., Praha: Státní nakladatelství technické literatury, 1986.
- [7] IRVING, C. a kol. *Twenty-year trends in prevalence and survival of Down syndrome*. *European Journal of Human Genetics*, 2008, 16.11: 1336-1340.
- [8] LOMTATIDZE, L., PLCH, R. *Sázíme v LaTeXu diplomovou práci z matematiky*. 1. vyd., Brno: Masarykova univerzita, 2003, 122 s., ISBN 80-210-3228-6.
- [9] MELOUN, M., MILITKÝ, J. *Statistická analýza experimentálních dat*. 2. vyd., up. a roz., Praha: Academia, 2004, 953 s., ISBN 80-200-1254-0.
- [10] MELOUN, M., MILITKÝ, J., HILL, M. *Statistická analýza vícerozměrných dat v příkladech*. 2. vyd., Praha: Academia, 2012, 750 s., Gerstner, sv. 7, ISBN 978-80-200-2071-0.
- [11] MORRIS, J. K. a kol. *Comparison of models of maternal age-specific risk for Down syndrome live births*. *Prenatal diagnosis*, 2003, 23.3: 252-258.

- [12] MORRIS, J. K., MUTTON, D. E., ALBERMAN, E. *Revised estimates of the maternal age specific live birth prevalence of Down's syndrome*. Journal of medical screening, 2002, 9.1: 2-6.
- [13] PAVLÍK, J. *Aplikovaná statistika*. Praha: Vysoká škola chemicko-technologická v Praze, 2005, ISBN 80-7080-569-2.
- [14] PROCHÁZKA, B. *Biostatistika pro lékaře: principy základních metod a jejich interpretace s využitím statistického systému R*. 1. vyd., Praha: Univerzita Karlova v Praze, nakladatelství Karolinum, 2015, ISBN 978-80-246-2782-3.
- [15] SAVVA, G. M. a kol. *Maternal age specific fetal loss rates in Down syndrome pregnancies*. Prenatal diagnosis, 2006, 26.6: 499-504.
- [16] SELIKOWITZ, M. *Downův syndrom: definice a příčiny, vývoj dítěte, výchova a vzdělání, dospělost*. 1. vyd., Praha. Portál, 2005.
- [17] SNIJDERS, R. JM. a kol. *Maternal age-and gestation-specific risk for trisomy 21*. Ultrasound in Obstetrics and Gynecology, 1999, 13.3: 167-170.
- [18] ŠMARDA, J. *Downův syndrom-jeho obraz z lékařského hlediska*. Časopis Speciální pedagogika, 2013, ISSN: 1211-2720.
- [19] Informační Bulletin České statistické společnosti [online], dostupné z: <http://www.statspol.cz/bulletiny/ib-2011-3-4-web.pdf>, [cit. 2015-11-14].
- [20] Labusová, E. *Metody prenatalní diagnostiky* [online], dostupné z: <http://www.evalabusova.cz/clanky/metody.php>, [cit. 2015-10-22].
- [21] Lifecodexx: Gutachten gegen Trisomie-21-Test [online], dostupné z: <https://www.biotechnologie.de/BIO/Navigation/DE/Aktuelles/politik,did=152972.html>, [cit. 2015-09-25].
- [22] Müller, I. *Časové řady*, Univerzita Palackého Olomouc, studijní opora [online], dostupné z: http://elearning-math.upol.cz/pluginfile.php/10846/mod_resource/content/2/03.pdf, [cit. 2016-03-28].
- [23] Společnost rodičů a přátel dětí s Downovým syndromem [online], dostupné z: <http://www.downsyndrom.cz/>, [cit. 2015-09-27].
- [24] Standardizace [online], dostupné z: http://www.demografie.info/?cz_umrtnoststandard=, [cit. 2016-03-16].