UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Gama rozdělení a jeho využití při analýze antimülleriánského hormonu

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **Mgr. Ondřej Vencálek Ph.D.**
Vypracoval: **Bc. Martin Vavruša**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2019

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Martin Vavruša

Název práce: Gama rozdělení a jeho využití při analýze antimülleriánského hormonu

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek Ph.D.

Rok obhajoby práce: 2019

Abstrakt: Antimülleriánský hormon je hormon v těle žen, který úzce souvisí s jejich možností otěhotnět. Tato diplomová práce se zaměřuje na to, jakým rozdělením pravděpodobnosti se tento hormon může řídit, a jak jeho hladina závisí na některých faktorech jako je například věk či kouření.

Klíčová slova: AMH, antimülleriánský hormon, gama rozdělení

Počet stran: 49

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Martin Vavruša

Title: Gamma distribution and its use in analysis of anti-Müllerian hormone

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek Ph.D.

The year of presentation: 2019

Abstract: Anti-Müllerian hormone is a hormone in a woman's body, which is closely related to her ability to become pregnant. This thesis is focused on looking for the correct probability distribution for this hormone and how it depends on covariates, such as age or smoking.

Key words: AMH, anti-Müllerian hormone, gamma distribution

Number of pages: 49

Number of appendices: 1

Language: Czech

Prohlášení

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

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Seznam tabulek

Poděkování

Děkuji Mgr. Ondřeji Vencálkovi Ph.D. za všechen čas, rady a připomínky, kterými přispěl k tvorbě této práce, a Fakultní nemocnici u sv. Anny v Brně, jež ochotně poskytla data ke zpracování praktické části.

Úvod

Cílem diplomové práce je modelovat hladinu antimülleriánského hormonu v závislosti na jiných vlastnostech ženského těla, na tom zda kouří a na přítomnosti některých chorob. Využívat budeme software R, v němž budeme analyzovat data z brněnské nemocnice sesbíraná v rozmezí let 2013 až 2017.

Na úvod čtenáře seznámíme s analyzovanými daty, včetně toho, co je antimülleriánský hormon a proč je předmětem zájmu. Ve druhé kapitole si nejprve představíme dva jednodušší modely předpokládající log-normální, resp. normální rozdělení tohoto hormonu, než se ve třetí kapitole pokusíme data modelovat s využitím gama rozdělení, u nějž odvodíme i některé jeho vlastnosti. Na závěr oba přístupy srovnáme a posoudíme, které rozdělení je pro tuto analýzu vhodnější.

Kapitola 1

Analyzovaná data

1.1 Antimülleriánský hormon

Předmětem našeho zájmu je antimülleriánský hormon (AMH), což je látka v těle žen, jejíž množství souvisí se schopností žen otěhotnět. Je produkován buňkami ve vaječnících a řídí proces dozrávání vajíček. Jeho hladina přímo souvisí s produkcí vajíček. Na základě množství tohoto hormonu v ženském těle se dá tedy poměrně přesně odhadnout počet vajíček a jakou má žena šanci na přirozené početí. U žen s nižší hladinou AMH bývají potraty častější, na základě znalosti hladiny tohoto hormonu v krvi se tedy může pacientka lépe připravit na případné potíže a začít uvažovat o možné léčbě či prevenci. Obecně se dá hodnota AMH interpretovat tak, jak je uvedeno v tabulce 1.1 (uvedené hodnoty jsou v $\mu g/l$).

Hodnota AMH	Hladina hormonu
$3 < AMH$	vysoká
$1 < AMH < 3$	normální
$0.7 < AMH < 0.9$	snížená
$0.3 < AMH < 0.7$	velmi snížená
$AMH < 0.3$	velmi nízká

Tabulka 1.1: Interpretace hladiny anti-mülleriánského hormonu[1]

1.2 Zkoumané regresory

Kromě hladiny AMH jsme měli u jednotlivých subjektů k dispozici dalších jedenáct informací. Datum narození a datum odběru jsme sloučili do proměnné věk, o jehož významném vlivu na hladinu AMH se již ví. Konkrétně od 23. roku života začíná hladina tohoto hormonu klesat a z původních 25 % klesne šance na přirozené početí na 20 % již do věku 30 let[1]. Kromě toho byl k dispozici údaj o roce, v němž k měření došlo (pracovali jsme s počtem let, které uplynuly od prvního roku měření, kterým byl rok 2013), a u většiny žen také údaje o výšce a váze, o délce menstruačního cyklu i samotné menstruace, množství testosteronu v těle, zda žena kouří či nikoliv, zda trpí konkrétními chorobami (PCOS, Amenorea)[1], a počet antralových folikulů (AFC), o němž byl však údaj dostupný jen pro měření v roce 2017, a z analýzy jsme jej tedy vyřadili. Následuje ukázka analyzovaných dat:

AMH	Výška	Váha	cyklus	PCOS	Test.	Kouří	Amen.	AFC	Rok
2.15	165	66	28/5	ne	1.48	ano	ne	NA	0.0
10.41	NA	NA	60/4-5	ne	2.10	ne	ne	NA	0.0
2.23	NA	NA	27/5	ne	1.12	ne	ne	NA	0.0
0.96	NA	NA	28/5	ne	NA	NA	ne	NA	0.0

Data bylo nejprve potřeba upravit, a to především kvůli poměrně velkému množství nulových hodnot u proměnné *hladina testosteronu*, pro kterou jsme zvolili následující kódování – ženy s hladinou testosteronu rovnou 0 $nmol/l$ jsme označili číslem 0, ženy s hladinou mezi 0 a 1.5 $nmol/l$ číslem 1 a ženy s hladinou testosteronu vyšší než 1.5 $nmol/l$ jsme zařadili do skupiny s označením 2. Konkrétní hladinu testosteronu jsme v další analýze neuvažovali. Podobně jsme ženy rozdělili do tří skupin ještě podle délky jejich menstruačního cyklu, konkrétně podle toho, jestli tato hodnota byla nižší než 27, v rozmezí 27 a 29, nebo vyšší než 29 dní. Délku samotné menstruace jsme poté zprůměrovali tam,

[1] *PCOS*, neboli *Syndrom polycystických ovárií*, je jedna z nejčastějších poruch žláz s vnitřní sekrecí u žen a je jednou z hlavních příčin jejich neplodnosti[2]. *Amenorea* je vynechání menstruačního krvácení u ženy v období pohlavní zralosti a plodnosti.[3]

kde nebylo uvedené konkrétní číslo, ale jen rozmezí. Výšku a váhu jsme naopak sloučili do jediné proměnné $BMI = \frac{hmotnost[kg]}{(výška[m])^2}$, abychom nemuseli uvažovat závislost mezi těmito dvěma proměnnými. Výsledná data, s nimiž jsme pracovali, tedy vypadala zhruba takto:

AMH	Mens. cyklus	Mens.	PCOS	Kouří	Věk	BMI	Test.	Amen.	Rok
2.15	0	5.0	ne	ano	19.31	24.24	1	ne	0.0
10.41	1	4.5	ne	ne	25.66	NA	2	ne	0.0
2.23	0	5.0	ne	ne	31.59	NA	1	ne	0.0
0.96	0	5.0	ne	NA	36.63	NA	NA	ne	0.0

Kapitola 2

Lineární regresní model

Lineární regresní model se zabývá úlohou modelovat střední hodnotu závisle proměnné náhodné veličiny y na nezávislých proměnných $\boldsymbol{x} = (x_1, \ldots, x_k)$, tzv. regresorech. Cílem modelu je odhadnout neznámé koeficienty $\beta_0, \ldots, \beta_k$, pro které platí:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

pro $i = 1, \ldots, n$, což lze přepsat také do maticového tvaru:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

kde

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$\boldsymbol{X}$ nazýváme datová nebo designová matice, $\boldsymbol{\beta}$ je vektor regresních koeficientů a $\boldsymbol{\epsilon}$ je vektor náhodných chyb.

2.1 Předpoklady lineárního regresního modelu

Lineární regresní model se opírá o čtyři základní předpoklady, které musejí být splněny, aby model správně fungoval:

- $n > k + 1$

- $h(\boldsymbol{X}) = k + 1$
- $E(\boldsymbol{\epsilon}) = \mathbf{0}$
- $var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$

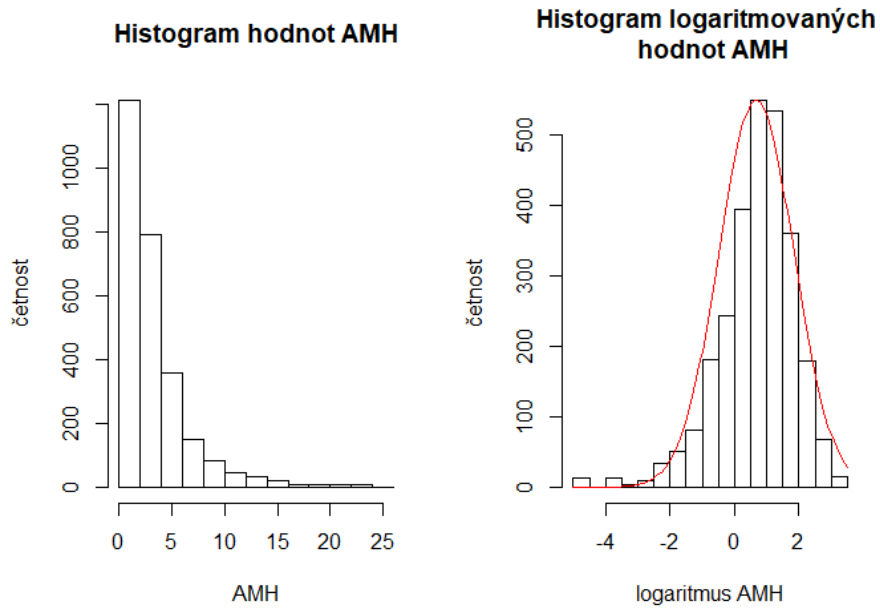
k zde značí počet vlastních regresorů v modelu, n počet pozorování, $\boldsymbol{X}$ matici regresorů hodnosti $h(\boldsymbol{X})$, $\boldsymbol{\epsilon}$ vektor chyb a $\boldsymbol{I}$ jednotkovou matici řádu n.

Za uvedených předpokladů můžeme metodou nejmenších čtverců získat nejlepší nestranný lineární odhad (NNLO) parametrů $\boldsymbol{\beta}$. Odvození jeho výpočtu zde nebudeme uvádět z důvodů rozsahu práce, zájemce můžeme odkázat například na publikaci Hron, Kunderová (2015)[4]. Zde uvedeme pouze konečný tvar odhadu, který je:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.$$

2.2 Rozdělení pravděpodobnosti závisle proměnné

První otázkou, se kterou se musíme potýkat, jestliže chceme statisticky modelovat nějakou veličinu, je, jaké rozdělení pravděpodobnosti můžeme u této náhodné veličiny předpokládat. V ideálním případě bychom chtěli, aby veličina měla normální rozdělení, pro které je statistická inference nejhlouběji prozkoumaná a výsledky jsou snáze interpretovatelné. K základnímu odhadu rozdělení pravděpodobnosti nám může posloužit histogram hodnot zkoumané veličiny.



Obrázek 2.1: Histogram hodnot AMH a logaritmovaných hodnot AMH

Z obrázku 2.1 je patrné, že normalita se v datech spíše očekávat nedá, čátečně také proto, že hladina hormonu AMH může nabývat pouze nezáporný hodnot, zatímco normální rozdělení nabývá s nenulovou pravděpodobností i hodnot záporných. Jestliže však tyto hodnoty nejprve zlogaritmujeme, a teprve poté se podíváme na jejich histogram, může se zdát, že se hladina hormonu řídí log-normálním rozdělením pravděpodobnosti, tedy že logaritmované hodnoty AMH pocházejí z normálního rozdělení. Odpovídající odhad křivky hustoty je v histogramu vykreslený červenou barvou. Hustotu log-normálního rozdělení lze poměrně jednoduše odvodit pouze ze znalosti hustoty normálního rozdělení:

Uvažujme náhodnou veličinu $X \sim N(\mu, \sigma^2)$, pro niž platí $X = ln(Y)$, resp. $Y = e^X$. Jednoduchým dosazením získáme distribuční funkci náhodné veličiny Y jako:

$$P(Y \leq y) = P(e^X \leq y) = P(X \leq ln(y)) = \Phi(ln(y)),$$

kde $\Phi(x)$ značí distribuční funkci náhodné veličiny X (tj. distribuční funkci normálního rozdělení). Její hustotu budeme značit $f(x)$. Hustotu náhodné veličiny Y s log-normálním rozdělením získáme derivováním složené funkce $\Phi(ln(y))$.

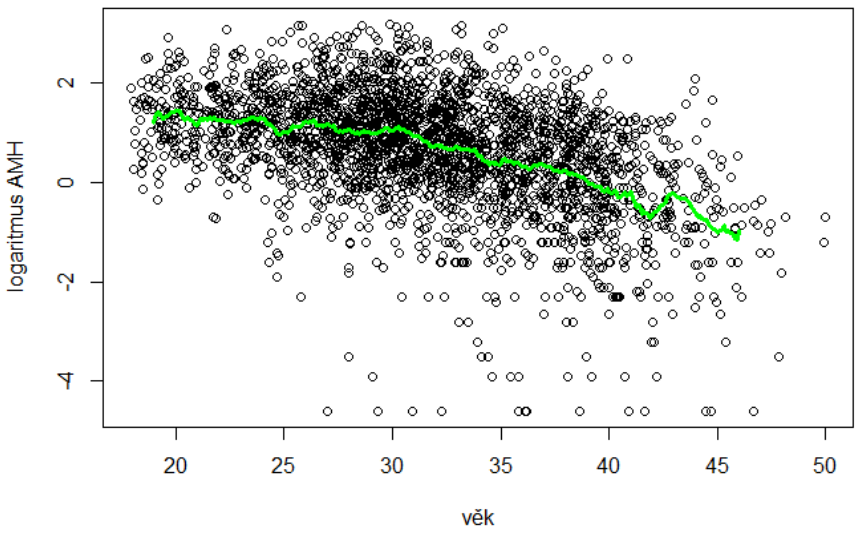
$$g(y) = \Phi'(ln(y)) = \frac{1}{y} f(ln(y)) = \frac{1}{y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(ln(y)-\mu)^2}{2\sigma^2}}. \tag{2.1}$$

V dalším textu tedy budeme jako závisle proměnnou uvažovat logaritmus hladiny AMH. Na konci druhé kapitoly a ve třetí kapitole, v níž se budeme věnovat zobecněným lineárním modelům a gama rozdělení, ale budeme využívat také původní hodnoty. Logaritmus hladiny AMH tedy budeme označovat symbolem $y^* = \ln(y)$.

2.3 Analýza dat o AMH pomocí lineárního regresního modelu

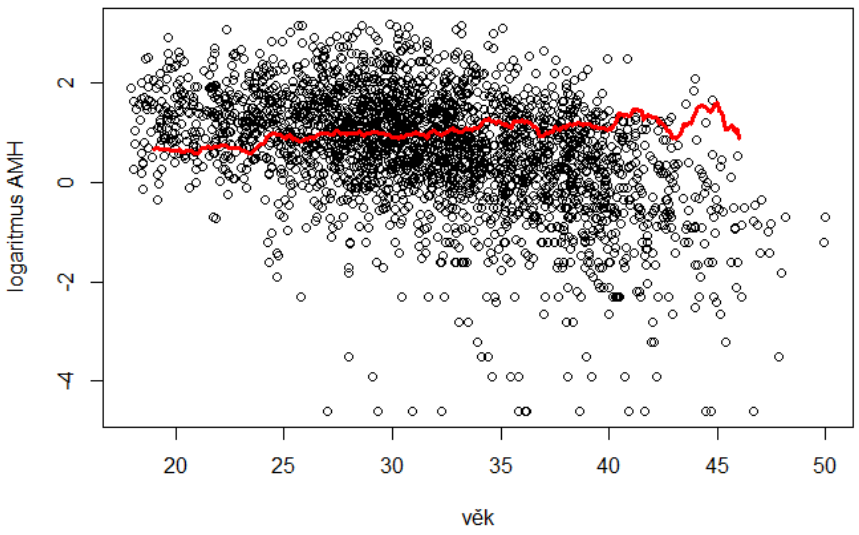
V prvním kroku analýzy budeme uvažovat jen závislost logaritmu AMH na věku x_i, pro níž budeme chtít znát přesnou podobu. Budeme tedy uvažovat závislost pouze dvou proměnných, kterou můžeme vykreslit do bodového grafu. Pro lepší představu o tom, jak se logaritmus AMH vyvíjí v průběhu stárnutí žen, vykreslíme do grafu také klouzavé průměry hodnot hladiny AMH, respektive jejich logaritmu, které nám zredukují informaci v datech. Klouzavé průměry pracují na následujícím principu:

Mějme n pozorování náhodné veličiny uspořádaných podle nějakého regresoru. V našem případě pozorování uspořádáme podle věku. Stanovíme hodnotu m, poloviny délky okna, například 0.6 let. První střed okna s tedy budeme uvažovat jako nejmenší hodnotu věku navýšenou o m. Vezmeme ženy s věkem v rozmezí $s - m$ až $s + m$ a spočítáme průměr z logaritmovaných hodnot jejich hladiny AMH. Poté s zvětšíme o nějakou vhodně zvolenou konstantu, třeba 0.1 a výpočet zopakujeme. Celý proces opakujeme, dokud $s+m$ neodpovídá maximální hodnotě věku. Výsledek je vyznačený zelenou barvou na obrázku 2.2.



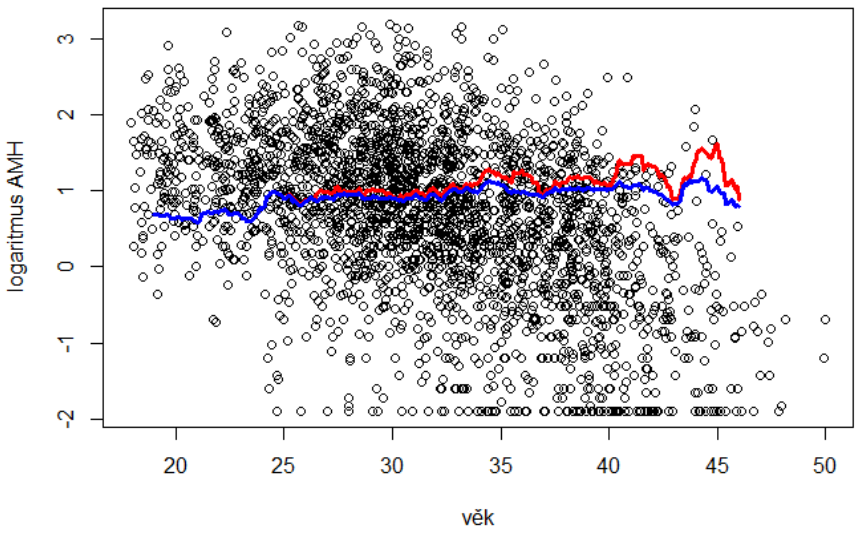
Obrázek 2.2: Klouzavé průměry počítané pro logaritmus hladiny AMH závisející na věku ženy

Budeme-li uvažovat jen regresor věku, je počet pozorování $n = 2\,726$ a počet vlastních regresorů $k = 1$. První podmínka je tedy splněna, a je splněna i druhá podmínka o hodnosti matice $\boldsymbol{X}$, $h(\boldsymbol{X}) = 2$. Ze způsobu odhadu parametrů $\boldsymbol{\beta}$ (metodou nejmenších čtverců) dostaneme i třetí podmínku a zbývá tedy ověřit jen předpoklad konstantního rozptylu náhodných odchylek pro všechna pozorování. Představu o splnění či nesplnění tohoto předpokladu můžeme získat, když sestavíme analogii klouzavých průměrů, kde namísto průměru budeme počítat směrodatnou odchylku z hodnot logaritmované hladiny hormonu v daném okně pro ženy ve věku od $s-m$ do $s+m$. Výsledek můžeme opět zakreslit do původního grafu závislosti mezi oběma proměnnými na obrázku 2.3. Vyznačený je červenou barvou.



Obrázek 2.3: Vývoj směrodatné odchylky logaritmu AMH v závislosti na věku žen

Z grafu je patrné, že hodnota směrodatné odchylky s rostoucím věkem roste, a čtvrtý předpoklad tedy není splněn. Navíc můžeme vidět, že z hlavního shluku pozorování některá výrazně vybočují, a proto nahradíme pozorování s hodnotou hladiny AMH nižší než 0.15, resp. v logaritmovaných datech ta, která jsou menší než -2.74 (celkem 75 hodnot, tj. asi 2.7 %), právě touto hraniční hodnotou. Tato odlehlá pozorování by totiž mohla být důsledkem chyby měření u velmi nízkých koncentrací. Směrodatná odchylka pak nebude v datech růst tak prudce, nicméně i nadále bude mít mírně rostoucí trend, jak můžeme vidět na obrázku 2.4.



Obrázek 2.4: Vývoj směrodatné odchylky logaritmu AMH v závislosti na věku žen před (červená křivka) a po odstranění potenciálně odlehlých pozorování (modrá křivka)

Abychom si s měnícím se rozptylem poradili, musíme model zobecnit a jeho parametry odhadovat metodou vážených nejmenších čtverců, kde za váhu v_i i-tého pozorování vezmeme druhou mocninu převrácené hodnoty věku x_i i-té ženy. Budeme tedy hledat koeficienty $\boldsymbol{\beta}$ minimalizující výraz

$$\sum_{i=1}^{n} \frac{1}{v_i}(y_i^* - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \frac{1}{x_i^2}(y_i^* - \beta_0 - \beta_1 x_i)^2,$$

což lze přepsat s pomocí vektoru $\boldsymbol{y^*}$ a matic $\boldsymbol{X}$ a $\boldsymbol{V}^{-1} = diag\{\frac{1}{x_1^2}, ..., \frac{1}{x_n^2}\}$ jako

$$(\boldsymbol{y^*} - \boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y^*} - \boldsymbol{X\beta}).$$

Tyto koeficienty můžeme nalézt tak, že výraz zderivujeme podle vektoru $\boldsymbol{\beta}$ a vektor těchto derivací položíme roven nulovému vektoru. Získáme tak soustavu rovnic

$$(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y^*},$$

z níž vyjádříme hledané koeficienty $\boldsymbol{\beta}$ jako:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}^*.$$

Odhad variační matice odhadu $\hat{\boldsymbol{\beta}}$ pak můžeme získat jako

$$\hat{var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}.$$

Detailní odvození vzorců lze najít např. ve Fišerová (2013)[5].

Pro takto spočítané hodnoty $\hat{\boldsymbol{\beta}}$ můžeme sestavit regresní přímku a vykreslit ji do pozorovaných dat. Je však potřeba mít na paměti, že jsme pracovali s logaritmovanými daty a k vyrovnaným hodnotám pro původní proměnnou AMH dospějeme pomocí vzorce

$$\hat{y}_i = e^{\hat{\beta}_0+\hat{\beta}_1 x_i+\frac{\hat{\sigma}^2 v_i}{2}}, \tag{2.2}$$

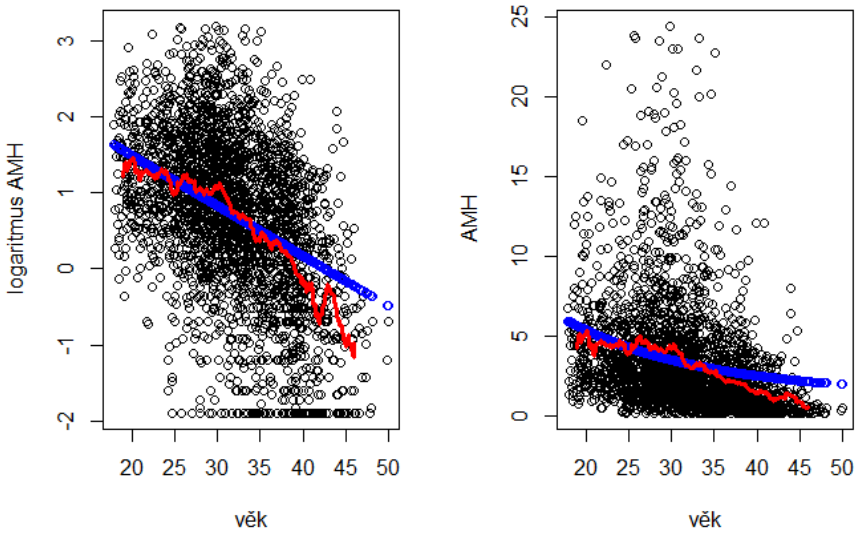
kde v_i je i-tý diagonální prvek matice $\boldsymbol{V}$. Vyrovnané hodnoty tedy spočítáme jako střední hodnotu odpovídajícího log-normálního rozdělení.

Graf vykreslíme jak pro logaritmovaná data, tak pro data původní. Výsledek je vidět na obrázku 2.5.

Odhad pomocí přímky zřejmě není nejvhodnější, proto se můžeme podívat ještě na to, jak by vypadala kvadratická či po částech lineární závislost. Výpočet vektoru koeficientů $\boldsymbol{\beta}$ je v takovém případě analogický předchozímu výpočtu. Stačí jen k regresoru x_i přidat regresor s hodnotami x_i^2 (v případě kvadratického trendu), nebo data rozdělit na dvě části a spočítat lineární trend pro každou z nich. Klíčovou roli v tomto případě bude hrát věk přibližně 30 let, který se ukázal jako nejvhodnější pro uvažování zlomu. Přesnější určení zlomového okamžiku a sestavení po částech lineárního modelu umožňuje v softwaru R příkaz *segmented()* v knihovně *segmented*. Výsledek je vidět na obrázku 2.6, kde můžeme nahoře vidět kvadratický trend a dole po částech lineární trend.

Oba tyto trendy už vypadají mnohem vhodněji než dříve uvedený lineární trend. Mezi nimi pak můžeme rozhodnout například na základě střední čtvercové chyby (anglicky *mean square error*).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

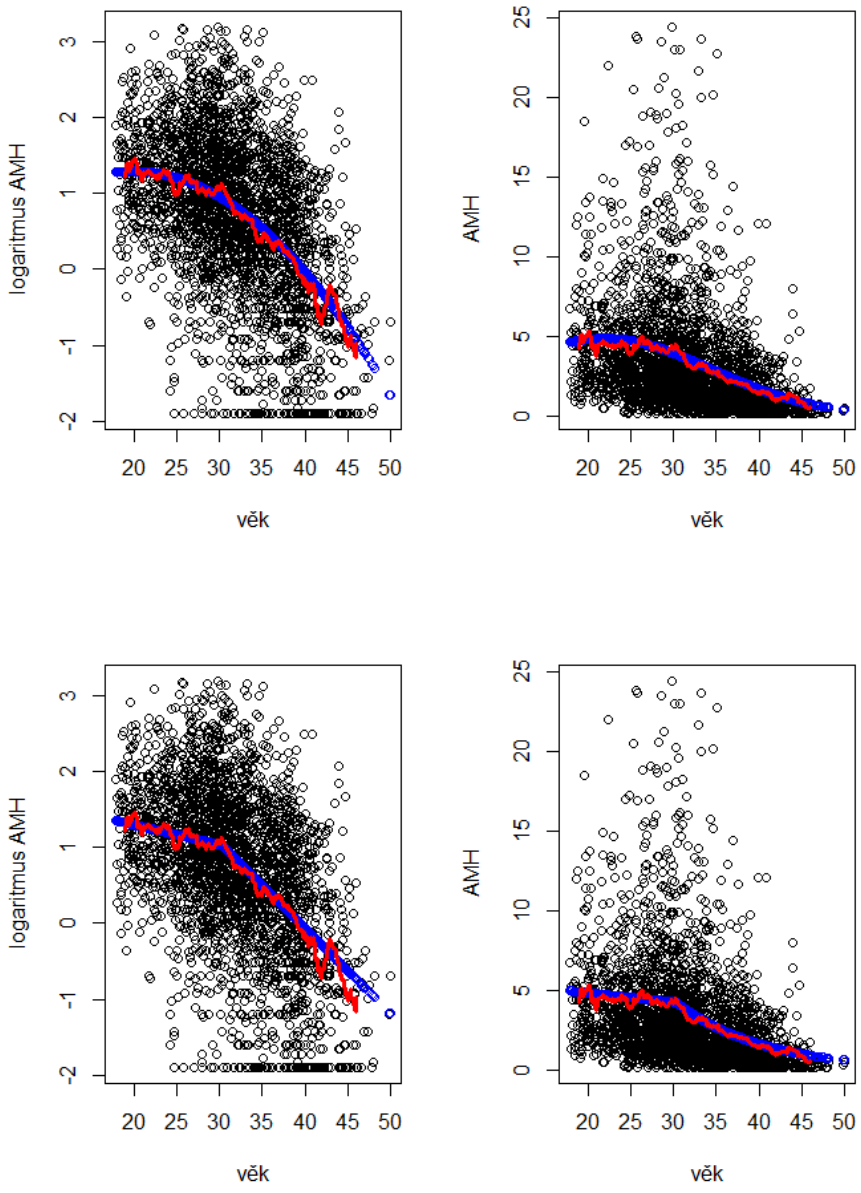


Obrázek 2.5: Závislost mezi AMH a věkem, včetně regresní křivky (modrá křivka) a klouzavých průměrů (červená křivka). Vlevo můžeme vidět logaritmovaná data, vpravo vidíme původní.

kde $\hat{y}_i$ je vyrovnaná hodnota získaná z rovnice 2.2 a y_i naměřená hodnota AMH pro i-tý subjekt. Toto kritérium obecně není příliš vhodné, ale s ohledem na to, že máme v obou modelech stejný počet parametrů, je v tomto případě přijatelné. Pro kvadratický trend nám vyjde hodnota MSE jako 11.26, pro po částech lineární trend vyjde podobná hodnota, 11.22. Na základě tohoto kritéria bychom tedy zvolili po částech lineární trend. K tomuto modelu následně můžeme zkusit přidat i další parametry, u nichž už budeme předpokládat, že na nich logaritmovaná hodnota AMH závisí lineárně.

K sestavení konečného modelu použijeme krokovou regresi (anglicky *stepwise regression*).

Kroková regrese spočívá v tom, že začneme od nejbohatšího, plného modelu, tj. od modelu, kde uvažujeme logaritmus AMH jako proměnnou závisející na všech uvedených regresorech, postupně odebereme každý regresor a posoudíme, který z nově vzniklých modelů je podle nějakého kritéria nejlepší, a zda je vůbec lepší než původní model. Jestliže bude nejlepší některý z modelů, kde chybí regresor,



Obrázek 2.6: Závislost mezi AMH a věkem, včetně regresní křivky. Vlevo vidíme logaritmovaná data (nahoře kvadratický trend, dole po částech lineární), vpravo je vidět výslednou regresní křivku pro původní data. Červená křivka představuje klouzavé průměry.

budeme tento dále uvažovat za plný model a celý výpočet zopakujeme. Pokud ne, budeme plný model považovat za správný. Postupovat se dá i opačným směrem, tj. k modelu pouze s absolutním členem přidáváme další regresory.

K posouzení dvou modelů existuje několik různých kritérií. My v této analýze využijeme *Akaikeho informační kritérium* (dále jen AIC), které je definováno jako

$$AIC = 2q - 2ln(\hat{L}),$$

kde q značí počet parametrů v daném modelu a $\hat{L}$ je funkce věrohodnosti daného modelu, která se spočítá jako

$$\hat{L} = \prod_{i=1}^{n} f(y_i^* | \boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$$

kde $f(y_i^* | \boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$ značí podmíněnou hustotu logaritmované hladiny AMH pro i-té pozorování při daných hodnotách regresorů $\boldsymbol{x}_i$ a koeficientů $\hat{\boldsymbol{\beta}}$. Pro takto spočítané AIC platí, že model s vyšší hodnotou tohoto kritéria datům odpovídá hůře než model s nižší hodnotou. Takto však můžeme porovnat jen modely vytvořené na stejném počtu pozorování, neboť množství dat má vliv na hodnotu $\hat{L}$.

Při práci v softwaru R můžeme tyto hodnoty získat snadno pomocí příkazu $AIC()$. Můžeme najít také příkaz $step()$, který přímo určí výsledný model získaný krokovou regresí, ten však vyžaduje jako argument datovou matici bez pozorování s chybějícími hodnotami, což může znatelně zredukovat informaci v datech. Proto se spokojíme s prvním uvedeným příkazem a budeme posuzovat dvojici modelů vždy na všech pozorováních, na kterých budeme umět vystavět oba porovnávané modely.

Provedeme-li výše uvedené analýzy na naší datové sadě, dostaneme, že logaritmus hladiny AMH závisí na věku, délce menstruačního cyklu (v tom smyslu, zda je průměrná, nadprůměrná nebo podprůměrná), PCOS, na tom, zda má žena hladinu testosteronu větší než 1.5 $nmol/l$, a také zjistíme, že mezi roky 2013 a 2017 průměrná hladina hormonu v populaci v čase stoupala. Závislost na věku

ženy přitom má podobu po částech lineární funkce, jak bylo uvedeno dříve. Ve výsledku dostaneme následující hodnoty odhadů parametrů β_j:

Regresor	$\hat{\beta}_j$	$\hat{SD}(\hat{\beta}_j)$	$e^{\hat{\beta}_j}$	p-hodnota
Absolutní člen	1.8505	0.1685	6.3632	< 0.0001
Věk (< 30 let)	-0.0396	0.0065	0.9612	< 0.0001
Věk (≥ 30 let)	-0.1116	0.0139	0.8944	< 0.0001
Menstruační cyklus	0.2085	0.0430	1.2318	< 0.0001
PCOS (přítomnost)	0.4314	0.0902	1.5394	< 0.0001
Testosteron (≥ 1.5)	0.1960	0.0431	1.2165	0.0141
Rok	0.0683	0.0155	1.0707	< 0.0001

Tabulka 2.1: Přehled parametrů pro lineární regresní model při modelování logaritmu AMH, včetně p-hodnot testů jejich významnosti

S ohledem na velké množství pozorování nesou p-hodnoty testů o parametrech jen jakousi doplňkovou informaci, neboť je známo, že s rostoucím počtem pozorování máme vyšší tendenci zamítat nulovou hypotézu ve prospěch alternativy.

2.3.1 Interpretace sestaveného modelu

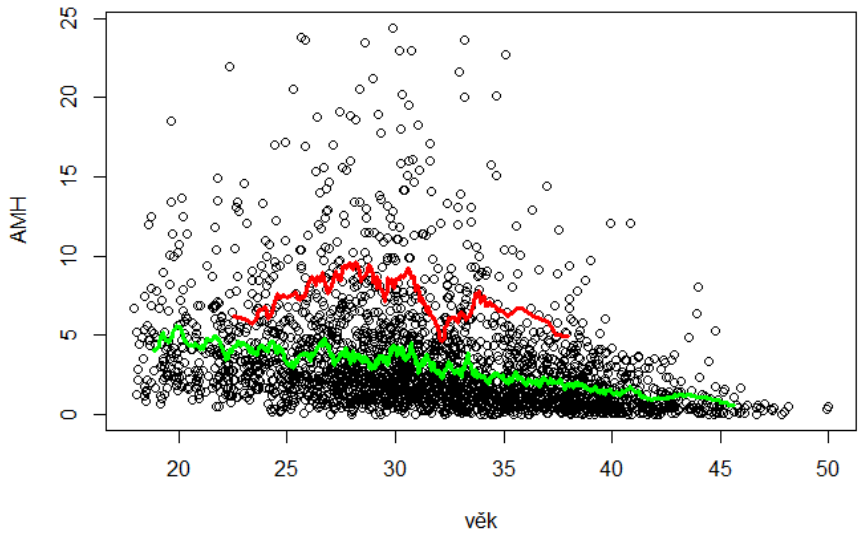
Při interpretaci si musíme nejprve uvědomit, že hodnoty hladiny AMH jsme logaritmovali před tím, než jsme s nimi jakkoliv pracovali, zatímco ostatní hodnoty jsme ponechali bez transformace. Změny v hladině hormonu AMH tak nebudeme dostávat v podobě absolutních čísel, ale jako procentuální změnu. Případně můžeme interpretovat hodnoty $e^{\hat{\beta}}$ jako koeficienty z multiplikativního modelu:

$$y_i = e^{\beta_0} e^{\beta_1 x_{i1}} \dots e^{\beta_k x_{ik}} e^{\epsilon}.$$

Absolutní člen se v dané situaci příliš interpretovat nedá, proto se jím zabývat nebudeme. Do třiceti let bychom řekli, že s každým dalším rokem života klesá v ženském těle hladina hormonu AMH v průměru o 4 % oproti roku předchozímu. Po třiceti letech už ale hladina klesá v průměru o 11 % za rok. Co se délky menstruačního cyklu týče, ženy s nadprůměrnou délkou tohoto cyklu mají přibližně o 20 % vyšší hladinu AMH, zatímco ženy s podprůměrnou délkou cyklu mají tento

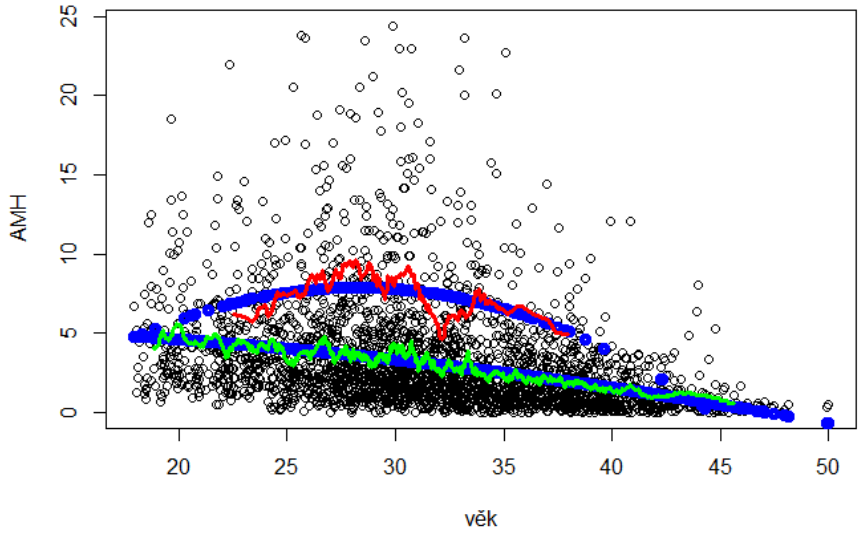
hormon v krvi zastoupen o 20 % méně. Ženy s koncentrací testosteronu větší než 1.5 $nmol/l$, mají rovněž v průměru 1.2 krát vyšší hladinu antimülleriánského hormonu oproti jiným ženám. Kromě toho nám vyšlo také, že mezi roky 2013 a 2017 stoupla průměrná hladina AMH v ženském těle každý rok přibližně o 7 %.

Interpretaci parametru u regresoru PCOS jsme schválně přeskočili, neboť ta si zaslouží větší pozornost. Sestavený model nám říká, že ženy s tímto syndromem mají v průměru o 43 % vyšší hladinu AMH (uvažujeme-li ostatní regresory neměnné). To může být dáno také tím, že by ženy s vyšší hladinou AMH měly větší tendenci trpět tímto syndromem. Podívejme se ale nejprve na klouzavé průměry AMH v závislosti na věku zvlášť pro ženy s PCOS (červeně) a bez něj (zeleně), aby naši představu nenarušovala závislost mezi jednotlivými koeficienty β_j. Vykreslíme-li graf do obrázku 2.7, nabízí se myšlenka sestavit lineární model na původních datech, který se bude lišit právě pro ženy s PCOS a bez něj, přičemž pro ženy bez tohoto syndromu bychom zřejmě uvažovali lineární závislost mezi věkem a hladinou hormonu v krvi.



Obrázek 2.7: Klouzavé průměry hladiny AMH v závislosti na věku zvlášť pro ženy bez PCOS (zeleně) a s PCOS (červeně)

Sestavíme tedy model s interakcí mezi věkem a PCOS, otázkou však zůstává, jakou závislost hladiny hormonu na věku můžeme očekávat u žen s tímto syndromem. Budeme-li ignorovat propad průměrné hladiny AMH u žen ve věku 30 až 35 let, které trpí syndromem PCOS, můžeme zkusit tuto závislost modelovat pomocí kvadratického trendu. Vyrovnané hodnoty modelu s interakcemi mezi PCOS a věkem, u něhož předpokládáme, že na něm hladina AMH závisí kvadraticky, můžeme vidět na následujícím obrázku.



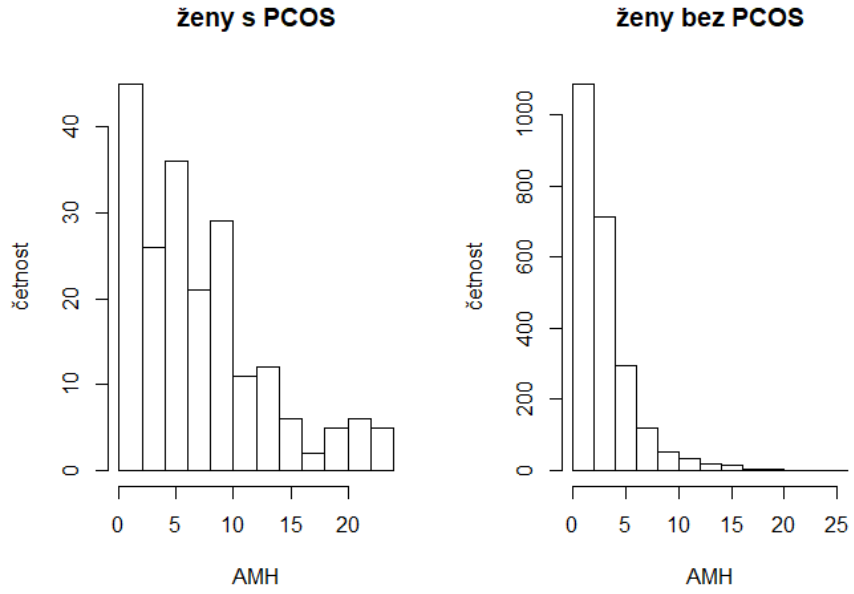
Obrázek 2.8: Klouzavé průměry hladiny AMH v závislosti na věku zvlášť pro ženy bez PCOS (zeleně) a s PCOS (červeně) spolu s vyrovnanými hodnotami modelu, který uvažuje interakce mezi PCOS a věkem (modrá křivka)

Stejně jako v předchozím případě najdeme pomocí krokové regrese také další významné regresory a dostaneme tak následující tabulku odhadů parametrů β_j a jejich vlastností (také tentokrát je potřeba mít na paměti, že p-hodnoty testů významnosti parametrů nesou jen doplňkovou informaci a s ohledem na velké množství dat budeme mít vyšší tendenci zamítat nulovou hypotézu):

Regresor	$\hat{\beta}_j$	$\hat{SD}(\hat{\beta}_j)$	p-hodnota
Absolutní člen	7.0065	1.4666	< 0.0001
Věk	-0.0624	0.0887	0.4819
Věk^2	-0.0015	0.0014	0.2711
Věk (při PCOS)	1.9407	0.3897	< 0.0001
Věk^2 (při PCOS)	-0.0304	0.0062	< 0.0001
Menstruační cyklus	0.6849	0.1279	< 0.0001
PCOS (přítomnost)	-27.2278	6.0589	< 0.0001
Testosteron (≥ 1.5)	0.6025	0.1352	< 0.0001
Amenorea	-1.1841	0.5493	0.0312
Rok	0.1990	0.0474	< 0.0001

Tabulka 2.2: Přehled odhadů parametrů pro lineární regresní model s interakcemi mezi PCOS a věkem při modelování hladiny AMH, včetně p-hodnot testů jejich významnosti

Zde však musíme ignorovat předpoklad normálního rozdělení závisle proměnné, což nám naznačuje už obrázek 2.1, kde na histogramu vidíme všechna data dohromady. Tento graf se nijak výrazně nepřiblíží histogramu pro normální rozdělení pravděpodobnosti, ani když data rozdělíme pro ženy s PCOS a bez něj, jak můžeme vidět na obrázku 2.9



Obrázek 2.9: Histogramy hladiny AMH pro ženy s PCOS a bez něj

Model, jehož regresní koeficienty jsme uvedli v tabulce 2.2, je v aditivním tvaru a uvedené koeficienty nám tedy značí přírůstek hladiny AMH při jednotkové změně příslušného regresoru, za předpokladu, že ostatní regresory zůstanou nezměněné. Například koeficient u délky menstruačního cyklu bychom interpretovali tak, že hladina AMH je v průměru o 0.65 $\mu g/l$ vyšší u žen s nadprůměrně dlouhým menstruačním cyklem než u žen s průměrnou délkou cyklu, které by se shodovaly ve všech ostatních hodnotách regresorů. Koeficienty u zbylých nezávisle proměnných bychom interpretovali podobně, jedinou výjimku tvoří regresor *věk*, který se s ohledem na nelineární podobu závislosti nedá jednoduše interpretovat. Maximalizací výrazu $-0.002x^2-0.062x$ resp. $-0.03x^2+1.941x$ (pro ženy s PCOS) však můžeme odhadnout věk ženy, do nějž bychom očekávali, že hodnota hladiny AMH v krvi poroste, a od nějž začne klesat. Pro ženy bez PCOS vychází řešení dané úlohy jako záporné číslo, po celou dobu bychom tedy očekávali, že hladina hormonu v krvi bude u těchto žen klesat, což koresponduje s obrázkem 2.8. Pro ženy se syndromem PCOS už ale vyjde hodnota kladná a na základě tohoto modelu bychom tvrdili, že pro tyto ženy roste hladina AMH v krvi přibližně do 32 let, a teprve od tohoto věku začíná klesat.

Kapitola 3

Zobecněný lineární model – gama rozdělení

3.1 Gama rozdělení

Toto rozdělení pravděpodobnosti spadá do širší rodiny exponenciálních rozdělení, jejichž hustota pro náhodnou veličinu Y je definována jako

$$f_Y(y;\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

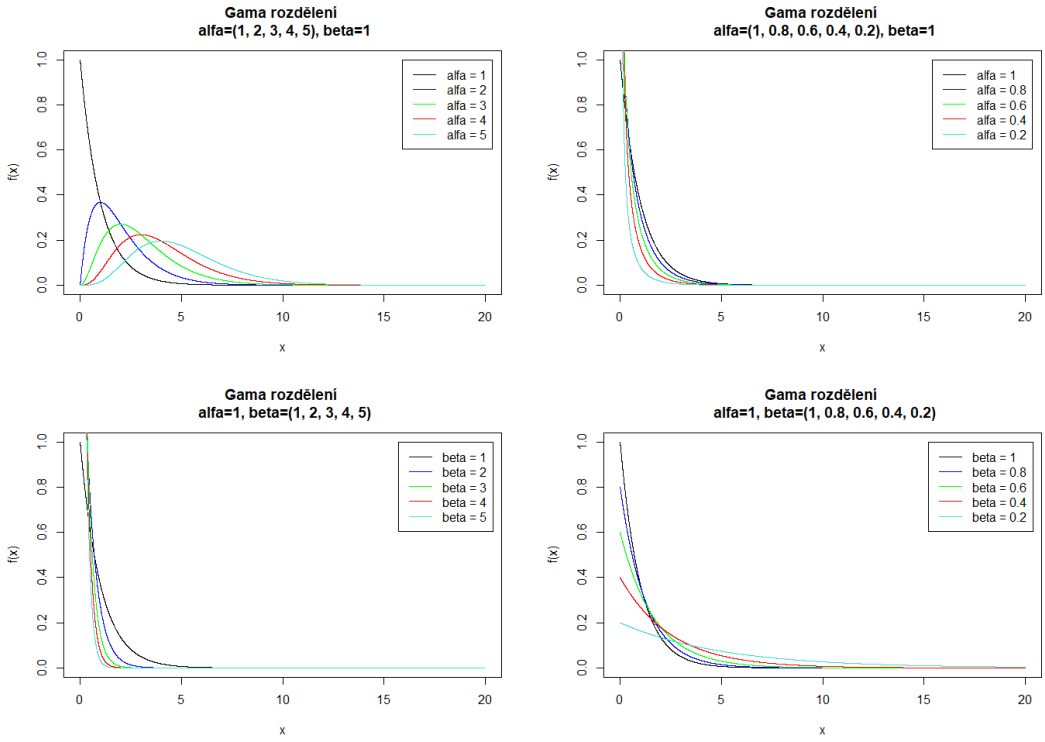
pro nějaké funkce a, b a c. Přímo gama rozdělení lze popsat s pomocí hustoty náhodné veličiny Y jako

$$f(y) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} & \text{pro } y > 0 \\ 0 & \text{pro } y \leq 0 \end{cases}, \tag{3.1}$$

kde $\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx$ je tzv. gama funkce. Abychom odlišili β ze vzorce pro hustotu gama rozdělení od regresních koeficientů, budeme je dále uvádět s dolním indexem Γ. Odvození hustoty gama rozdělení je popsáno v McCullagh, Nelder (1989)[6].

Gama rozdělení má oproti lognormálnímu rozdělení výhodu v tom, že je mnohem flexibilnější a může nabývat nejrůznějších tvarů. Některé jeho možné podoby jsou uvedeny na obrázku 3.1.

Parametr α tohoto rozdělení se nazývá parametr tvaru, β_Γ je inverzní parametr k parametru měřítka. Pro konkrétní data můžeme odhadnout tyto parame-



Obrázek 3.1: Některé podoby hustoty gama rozdělení – nahoře při pevně daném $\beta_\Gamma = 1$ a měnícím se parametru α, dole při daném $\alpha = 1$ a měnícím se parametru β_Γ.

try buď numericky, maximalizováním funkce věrohodnosti za podmínek $\alpha > 0$ a $\beta_\Gamma > 0$, nebo z bodových odhadů střední hodnoty a rozptylu. Platí totiž:

$$E(Y) = \mu = \frac{\alpha}{\beta_\Gamma} \qquad var(Y) = \frac{\alpha}{\beta_\Gamma^2}. \tag{3.2}$$

Maximálně věrohodný odhad střední hodnoty můžeme získat jednoduše pomocí aritmetického průměru, pro rozptyl je však vzorec komplikovanější a využívá funkci, která se nazývá *deviance*. V případě gama rozdělení ji můžeme spočítat jako

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2 \sum_i w_i \left(\ln \frac{y_i}{\hat{\mu}_i} - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right),$$

kde w_i jsou známé váhy rozptylu jednotlivých pozorování, podobně jako jsme s nimi pracovali u lineárního regresního modelu. Odvození tohoto výpočtu lze najít v publikaci McCullagh a Nelder (1989)[6]. Devianci ale lze spočítat jen za předpokladu, že jsou všechny pozorované hodnoty kladné, a je tedy potřeba dát si pozor na nulové hodnoty, které mohou vzniknout zaokrouhlováním.

Jestliže známe devianci, získáme maximálně věrohodný odhad parametru α z rovnice:

$$\frac{1}{\hat{\alpha}} \approx \frac{\bar{D}(6 + \bar{D})}{6 + 2\bar{D}}, \tag{3.3}$$

kde

$$\bar{D} = \frac{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}{n}.$$

V softwaru R můžeme tuto hodnotu získat s pomocí funkce *gamma.dispersion()* z knihovny *MASS*. Z odhadu $\frac{1}{\hat{\alpha}}$ získáme odhad rozptylu náhodné veličiny jako

$$\hat{var}(Y) = \frac{1}{\hat{\alpha}} \hat{\mu}^2.$$

3.2 Zobecněný lineární model

Ve druhé kapitole jsme se věnovali lineárnímu regresnímu modelu, který můžeme vyjádřit ve tvaru $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Jinými slovy, závisle proměnnou $\boldsymbol{y}$ jsme

vyjádřili jako součet její systematické části $\boldsymbol{\mu} = \boldsymbol{X\beta}$ a náhodné složky $\boldsymbol{\epsilon}$. Zobecnění lineárního regresního modelu spočívá v tom, že systematickou část $\boldsymbol{\eta} = \boldsymbol{X\beta}$ nebudeme uvažovat přímo jako $\boldsymbol{\mu}$, ale jako jeho funkci $\eta_i = g(\mu_i), i = 1, \ldots, n$ kde μ_i a η_i představují i-tou složku vektoru $\boldsymbol{\mu}$, resp. $\boldsymbol{\eta}$. Funkci g nazýváme tzv. *link function*.

V případě normálního rozdělení bychom dostali $\mu_i = \eta_i, i = 1, \ldots, n$. Při gama rozdělení závisle proměnné bude mít tato funkce tvar

$$\eta_i = \mu_i^{-1} \qquad i = 1, \ldots, n.$$

Odhady parametrů lze odvodit analyticky pouze pro některé jednoduché případy, jako je model, kde

$$\eta = \beta_0 + \beta_1 x,$$

přičemž $x \in \{0, 1\}$. V takovém případě můžeme získat odhady parametrů β_0 a β_1 tak, že budeme maximalizovat logaritmus věrohodnostní funkce. Ten v případě gama rozdělení a pro jediné pozorování y vypadá následovně:

$$L(y; \alpha, \mu) = \alpha\Big(-\frac{y}{\mu} - ln(\mu)\Big) + \alpha ln(y) + \alpha ln(\alpha) - ln\Big(\Gamma(\alpha)\Big). \tag{3.4}$$

Tento výraz získáme, když za parametr β_Γ v předpisu hustoty gama rozdělení 3.1 dosadíme $\beta_\Gamma = \frac{\alpha}{\mu}$ (tedy odvodíme je ze vztahu pro střední hodnotu Y), a následně zlogaritmujeme. Tuto věrohodnostní funkci chceme maximalizovat přes parametry β_0 a β_1, na nichž v uvedeném předpisu zřejmě závisí parametr μ. Budeme předpokládat, že parametr α na regresních koeficientech nezávisí. Ostatní členy potom můžeme považovat za konstanty a maximalizovat pouze výraz

$$-\frac{y}{\mu} - ln(\mu).$$

Po vynásobení mínus jedničkou a pro soubor n pozorování tak dostaneme úlohu minimalizovat

$$\sum_{i=1}^{n} \Big[\frac{y_i}{\mu_i} + ln(\mu_i)\Big].$$

Nyní už stačí jen dosadit za $\mu_i = \frac{1}{\beta_0+\beta_1 x_i}$, zderivovat výraz podle obou parametrů $\beta_j, j = 0, 1$, a tyto derivace srovnat s nulou. Výsledkem bude soustava rovnic

$$\sum_i \left(y_i - \frac{1}{\beta_0 + \beta_1 x_i} \right) = 0$$

$$\sum_i \left(y_i x_i - \frac{x_i}{\beta_0 + \beta_1 x_i} \right) = 0,$$

jejímž řešením jsou maximálně věrohodné odhady regresních parametrů

$$\hat{\beta}_0 = \frac{1}{\frac{1}{\#i:x_i=0} \sum_{i:x_i=0} y_i} \qquad \hat{\beta}_1 = \frac{1}{\frac{1}{\#i:x_i=1} \sum_{i:x_i=1} y_i} - \hat{\beta}_0.$$

V ostatních případech odhad parametrů není zdaleka tak jednoduchý a musíme využít některou z iteračních metod. Příkladem takové metody může být IRWLS algoritmus (*iterative reweighted least squares*[7]):

1. Zvolíme počáteční hodnoty $\hat{\boldsymbol{\beta}}^{(0)}$ a $\hat{\mu}^{(0)}$

2. V k-tém kroku spočítáme upravenou závisle proměnnou

   $$z_i^{(k)} = \boldsymbol{x_{i\cdot}}' \hat{\boldsymbol{\beta}}^{(k-1)} + (y_i - \hat{\mu}_i^{(k-1)}) \frac{\partial g(\mu)}{\partial \mu}|_{\mu=\hat{\mu}_i^{(k-1)}} \qquad i = 1, \ldots, n,$$

   kde $g(\mu) = \frac{1}{\mu}$ je link function a $\boldsymbol{x_{i\cdot}}$ je sloupcový vektor tvořený hodnotami i-tého řádku matice $\boldsymbol{X}$.

3. Spočítáme vektor vah

   $$\left(w_i^{(k)}\right)^{-1} = V(\hat{\mu}_i^{(k-1)}) \left(\frac{\partial g(\mu)}{\partial \mu}|_{\mu=\hat{\mu}_i^{(k-1)}}\right)^2 = (\hat{\mu}_i{}^{(k-1)})^{-2} \qquad i = 1, \ldots, n,$$

   neboť v případě gama rozdělení náhodné veličiny Y je $V(\mu_i) = \mu_i^2$ (viz McCullagh, Nelder[6]) a derivací $g(\mu)$ získáme výraz $-\frac{1}{\mu^2}$

4. Aktualizujeme $\hat{\boldsymbol{\beta}}^{(k-1)}$ na $\hat{\boldsymbol{\beta}}^{(k)}$ pomocí metody vážených nejmenších čtverců, kde závisle proměnnou je $\boldsymbol{z}^{(k)} = (z_1^{(k)}, \ldots, z_n^{(k)})$, nezávisle proměnné jsou obsaženy v matici $\boldsymbol{X}$ a váhy představuje vektor $\boldsymbol{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})$

Další informací, kterou budeme chtít o odhadech parametrů $\boldsymbol{\beta}$ znát, je jejich variabilita. Tu můžeme odhadnout jako

$$\hat{var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X'}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1},$$

kde $\hat{\boldsymbol{W}}$ je odhad matice

$$\boldsymbol{W} = diag\left\{\frac{\left(\frac{d\mu_i}{d\eta_i}\right)^2}{V(\mu_i)}\right\}.$$

Jak jsme uvedli výše, při gama rozdělení závisle proměnné platí $V(\mu_i) = \mu_i^2$. Navíc $\frac{d\mu_i}{d\eta_i} = \frac{\partial g^{-1}(\eta)}{\partial \eta}|_{\eta=\eta_i} = -\frac{1}{\eta_i^2} = -\mu_i^2$ a matici $\boldsymbol{W}$ můžeme odhadnout jako

$$\hat{\boldsymbol{W}} = diag\left\{\hat{\mu}_i^2\right\}.$$

Maximálně věrohodný odhad rozptylu jsme uvedli výše, nicméně chceme-li konzistentní odhady regresních parametrů $\boldsymbol{\beta}$, musíme použít jiný vzorec pro odhad $\frac{1}{\alpha}$, který vede ke konzistentnímu odhadu rozptylu:

$$\frac{1}{\hat{\alpha}} = \sum_i \frac{(\frac{y_i - \hat{\mu_i}}{\hat{\mu_i}})^2}{(n-p)}.$$

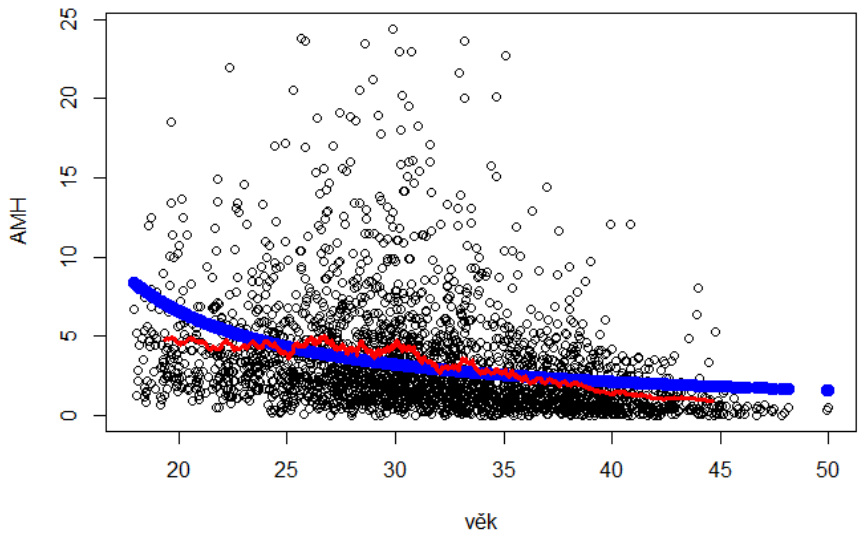
Ten můžeme v softwaru R získat pomocí funkce *summary()*, do níž dosadíme objekt *glm*. Stejně tak nám tato funkce dá odpověď i na otázku, jaká je směrodatná odchylka pro odhady jednotlivých regresních koeficientů β_j.

3.3 Analýza dat o AMH pomocí zobecněného lineárního modelu

Analýzu dat opět zahájíme výběrem vhodného modelu závislosti AMH na věku. Tak jako v předchozím případě začneme tím, že prozkoumáme, jak dobře by data proložil jednoduchý zobecněný lineární model, který uvažuje regresor věku bez jakékoliv transformace. Systematická část modelu tedy bude vypadat následovně:

$$\eta = \beta_0 + \beta_1 x,$$

kde x značí věk ženy.



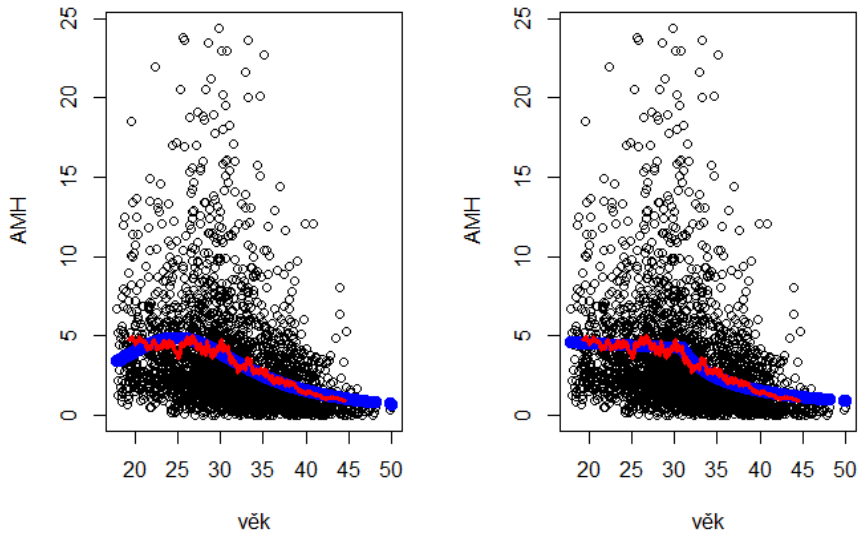
Obrázek 3.2: Srovnání jednoduchého zobecněného lineárního modelu závislosti AMH na věku a klouzavých průměrů na zkoumaných datech

Vykreslíme-li vyrovnané hodnoty do grafu na obrázku 3.2 (modře) a srovnáme je s klouzavými průměry (červená křivka), uvidíme, že ani v případě zobecněných lineárních modelů nestačí uvažovat věk jako regresor bez jakékoliv transformace. Můžeme proto opět zkusit uvažovat kvadratický nebo po částech lineární trend hladiny AMH v závislosti na věku.

$$\eta = \beta_0 + \beta_1 x + \beta_2 x^2$$

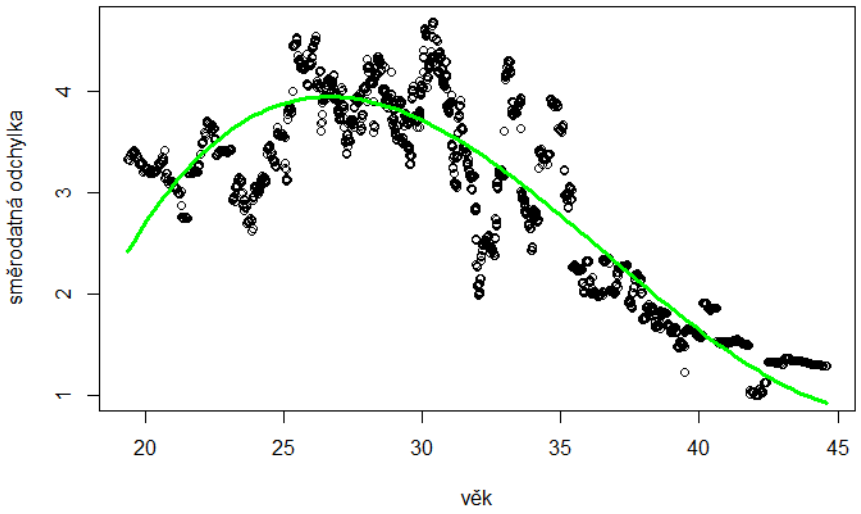
$$\eta = \beta_0 + \beta_1 x \cdot I_{\left[x<30\right]} + \beta_2 x \cdot I_{\left[x\geq 30\right]}$$

Jak moc budou vyrovnané hodnoty obou modelů odpovídat klouzavým průměrům, můžeme posoudit nejprve na základě grafů na obrázku 3.3.



Obrázek 3.3: Srovnání kvadratického (vlevo) a po částech lineárního trendu (vpravo) při modelování závislosti AMH na věku pomocí zobecněného lineárního modelu spolu s klouzavými průměry na zkoumaných datech

Oba odhady jsou od prvního pohledu podstatně lepší než model lineární závislosti AMH na věku a přinejmenším velmi podobné dříve uvažovanému modelu s předpokladem log-normálního rozdělení po přidání vah. Ty do modelu můžeme samozřejmě zahrnout i v případě zobecněného lineárního modelu. Jejich tvar můžeme odvodit tak jako u lineárního regresního modelu, vykreslením analogie klouzavých průměrů s využitím směrodatné odchylky. Tentokrát už však budeme pracovat s původními hodnotami AMH, namísto s jejich transformací pomocí přirozeného logaritmu. Odhad trendu ve směrodatné odchylce při dané délce okna $m = 100$ pozorování (uspořádaných podle věku pacientky) a posunu vždy o jedno pozorování, můžeme vidět na obrázku 3.4. Do grafu je vykreslený také odhad kubického trendu získaný metodou nejmenších čtverců.

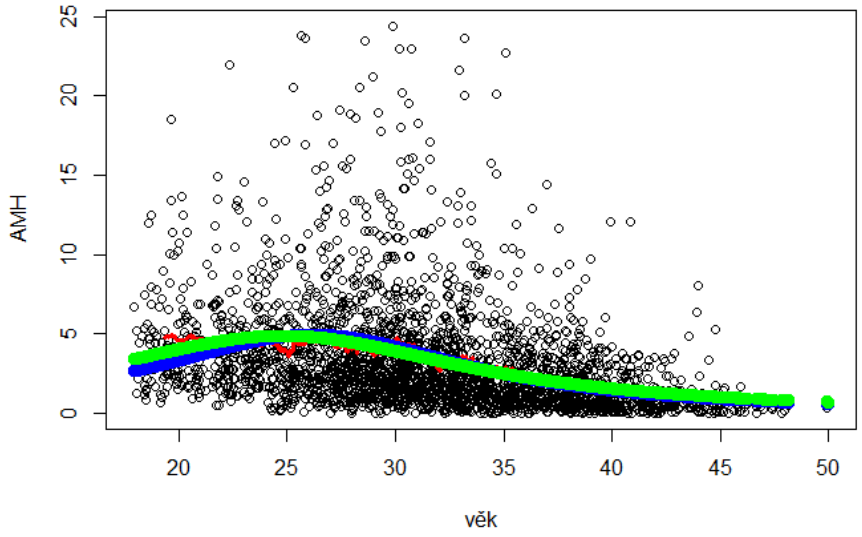


Obrázek 3.4: Odhad směrodatné odchylky v datech o AMH spolu s odhadem kubického trendu (zelená křivka)

Druhou cestou, jak odhadnout trend ve směrodatné odchylce, by bylo využít znalostí z procesu tvorby lineárního regresního modelu, kde jsme odhadovali trend ve střední hodnotě i ve směrodatné odchylce pro logaritmovaná data. Označíme-li tyto trendy jako $\mu_L(x)$ a $\sigma_L(x)$, kde x značí věk ženy, můžeme ze vztahů pro výpočet rozptylu log-normálního rozdělení spočítat odhad směrodatné odchylky v původních datech jako

$$\sigma(x) = \sqrt{(e^{\sigma_L^2} - 1)e^{2\mu_L + \sigma_L^2}}.$$

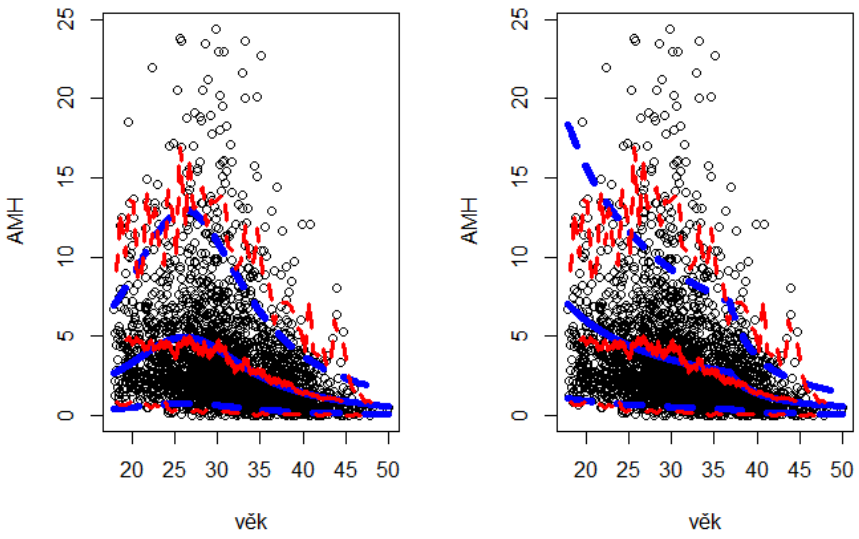
Odhad kubickým trendem se zdá poměrně přijatelný, nicméně zahrneme-li váhy do zobecněného lineárního modelu, docílíme jen nepatrné změny. Oproti tomu, jestliže se podíváme na totéž srovnání také při použití vah odvozených z odhadů u log-normálního rozdělení pravděpodobnosti, budou se výsledky lišit podstatně více, jak je vidět na obrázku 3.5. Váhy je tedy vhodné do modelu zahrnout, nicméně kubický trend směrodatnou odhylku neodhaduje zdaleka dobře. U po částech lineárního trendu bychom došli k podobnému závěru.



Obrázek 3.5: Srovnání zobecněného lineárního modelu (zelená křivka) a váženého zobecněného lineárního modelu (modrá křivka) při kvadratické závislosti AMH na věku a vahách odhadnutých na logaritmovaných datech

Kdybychom se chtěli rozhodnout mezi uvedenými dvěma typy závislosti AMH na věku, můžeme opět využít některé z obecně známých kritérií, ať už uvedené AIC, MSE, nebo například průměrnou absolutní chybu predikce (MAPE). My se ale podíváme ještě na jeden způsob, jak bychom mohli vybrat model na základě grafu, a sice vykreslíme kromě klouzavých průměrů také nějaký odhad 2.5% a 97.5% kvantilu gama rozdělení v závislosti na věku. To můžeme udělat tak, že data rozdělíme do skupin podle stáří žen, přičemž věkový rozdíl mezi ženami v rámci jedné skupiny nebude větší než 0.5 let. Pro každou skupinu pozorování pak odhadneme parametry α a β_Γ momentovou metodou, tj. s pomocí vztahů 3.2. Tyto kvantily pak budeme porovnávat s kvantily gama rozdělení pro každou vyrovnanou hodnotu z modelu s kvadratickou resp. po částech lineární závislostí AMH na věku. Jako střední hodnotu jednotlivých rozdělení pravděpodobnosti budeme uvažovat vyrovnanou hodnotu $\hat{y}_i$ z jednoho nebo druhého modelu, a rozptyl získáme jako $\frac{1}{\hat{\alpha}_i}\hat{y}_i^2$. Výsledné grafy vidíme na obrázku 3.6. Červenou křivkou jsou vykresleny klouzavé průměry hladiny AMH v závislosti na věku a přerušovanými

čarami také kvantily gama rozdělení pro data rozdělená do menších skupin. Modře je odhadnutá křivka závislosti AMH na věku (vlevo kvadratická závislost, vpravo po částech lineární závislost), přerušovaně pak příslušné kvantily gama rozdělení.



Obrázek 3.6: Srovnání modelů uvažujících různou závislost AMH na věku a odhadnutých kvantilů rozdělení pravděpodobnosti pro hladinu hormonu v krvi při daném věku (vlevo model s kvadratickou závislostí AMH na věku, vpravo model uvažující po částech lineární závislost)

Na základě tohoto srovnání můžeme vidět, že model předpokládající kvadratický trend lépe popisuje variabilitu AMH, a do další analýzy tedy budeme uvažovat kvadratickou závislost tohoto hormonu na věku. Ovšem v případě, že bychom uvažovali po částech lineární závislost, dostaneme tytéž významné regresory jako v případě kvadratického trendu. Odhady regresních koeficientů se budou lišit v řádech setin.

K sestavení konečného modelu využijeme také tentokrát krokovou regresi, kterou jsme popsali již ve druhé kapitole. Uvedeme proto jen konečný výčet významných regresorů a odhady příslušných regresních koeficientů. Jejich přehled doplněný o p-hodnotu testu významnosti pro jednotlivé parametry v modelu je uveden v tabulce 3.1 (uvedené jsou konzistentní odhady směrodatných odchylek).

Regresor	$\hat{\beta}_i$	$\hat{SD}(\hat{\beta}_i)$	p-hodnota
Absolutní člen	1.8900	0.2085	< 0.0001
Věk	-0.1230	0.0134	< 0.0001
Věk^2	0.0024	0.0002	< 0.0001
Menstruační cyklus	-0.0832	0.0152	< 0.0001
PCOS (přítomnost)	-0.0899	0.0192	< 0.0001
Rok	-0.0199	0.0057	0.0005

Tabulka 3.1: Přehled parametrů pro vážený zobecněný lineární model s předpokladem gama rozdělení hladiny AMH a kvadratické závislosti této proměnné na věku, včetně p-hodnot testů jejich významnosti

Zde narazíme na problém s interpretací regresních koeficientů $\boldsymbol{\beta}$, které s ohledem na tvar modelu nelze interpretovat tak, jako tomu bylo u modelů probíraných ve druhé kapitole. Můžeme z nich vyčíst pouze to, jestli jednotlivé regresory působí na hladinu AMH v pozitivním nebo negativním smyslu. Pozitivní vliv budou mít regresory, u nichž jsme odhadli parametr β_j záporně, neboť s roustoucím x bude klesat převrácená hodnota střední hodnoty gama rozdělení, tj. samotná střední hodnota bude růst. Podobně můžeme říci, že negativní vliv mají regresory, pro něž jsme parametr β_j odhadli kladně. Vidíme tedy, že nadprůměrná délka menstruačního cyklu a PCOS zvyšují hladinu AMH, stejně jako to, že mezi roky 2013 a 2017 průměrná hladina AMH u žen rostla. K tomu, abychom mohli rozhodnout o vlivu věku, musíme najít stacionární bod funkce

$$f(x) = \frac{1}{0.0024x^2 - 0.1230x + c},$$

kde $c \neq -0.0024x^2 + 0.1230x$ je konstanta zahrnující pevně dané hodnoty ostatních regresorů. Nalezený stacionární bod 25.625 je s ohledem na tvar funkce bodem jejího maxima, a udává nám, do kterého roku života AMH v závislosti na věku žen roste, a od jakého věku klesá.

Kapitola 4

Srovnání modelů pro analýzu AMH

Dosud jsme srovnávali jen modely se stejnými předpoklady (např. jsme mezi sebou srovnávali jen modely předpokládající log-normální rozdělení) a různými regresory, tedy hledali jsme nejlepší model za určitých předpokladů. Dospěli jsme tak ke třem různým modelům, které bychom potřebovali porovnat také vzájemně mezi sebou. AIC v tomto případě použít nemůžeme, jelikož hned dva modely pracují s vahami, a ty by se projevily také v hodnotě tohoto kritéria. K porovnání modelů tak použijeme cross-validaci.

Cross-validace modelu spočívá v tom, že z dat vynecháme skupinu pozorování (nebo jako v našem případě jediné pozorování), na ostatních odhadneme model a zkusíme s jeho pomocí predikovat pozorování, která jsme vynechali. To potom můžeme porovnat se skutečnými naměřenými hodnotami a zjistit velikost chyb, tj. absolutní odchylky skutečných a predikovaných hodnot. Celý proces opakujeme dokud postupně nevynecháme všechna pozorování.

Celkem tedy dostaneme soubor téměř 2000 hodnot pro každý model (po vynechání chybějících hodnot), a tato data můžeme podrobit další analýze.

Pro připomenutí, modely, které jsme vybrali jako nejlepší za určitých předpokladů, jsou tyto:

1. Model 1: Model předpokládající podmíněné log-normální rozdělení pravděpodobnosti závisle proměnné, přičemž závislost AMH na věku má podobu

	model 1	model 2	model 3
Minimum	0.0008	0.0035	0.0007
1. kvartil	0.4694	0.6687	0.6735
Medián	1.0235	1.3834	1.3303
Průměr	1.7906	1.8717	1.9463
3. kvantil	2.0622	2.2842	2.3317
Maximum	20.4482	19.6937	19.3337
Rozptyl	5.6090	4.3040	5.0573

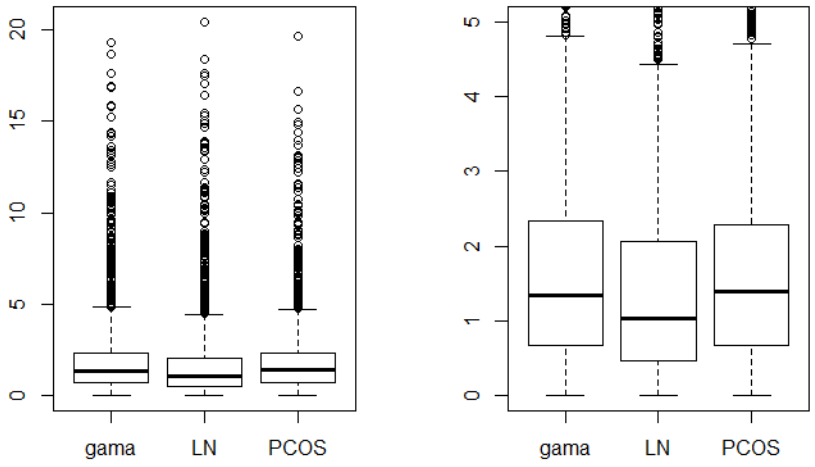
Tabulka 4.1: Shrnutí chyb cross-validace pro jednotlivé použité modely

po částech lineární funkce a další regresory jsou délka menstruačního cyklu, PCOS, zda je hladina testosteronu vyšší než 1.5 $nmol/l$ a rok, v němž k měření došlo.

2. Model 2: Lineární regresní model s interakcí mezi věkem a PCOS. Závislost mezi hladinou AMH a věkem uvažujeme kvadratickou. Další regresory jsou délka menstruačního cyklu, PCOS, zda je hladina testosteronu vyšší než 1.5 $nmol/l$, rok, v němž k měření došlo a amenorea.

3. Model 3: Zobecněný lineární model, který předpokládá podmíněné gama rozdělení závisle proměnné, kvadratickou závislost AMH na věku, a jehož dalšími regresory jsou délka menstruačního cyklu, PCOS a rok měření.

Jednou z možností, jak vyhodnotit chyby získané cross-validací a porovnat na jejich základě uvedené modely, je podívat se na důležité číselné charakteristiky, jejichž hodnoty můžeme shrnout do tabulky 4.1. Případně můžeme některé charakteristiky vykreslit také pomocí boxplotu, který je vidět na obrázku 4.1.

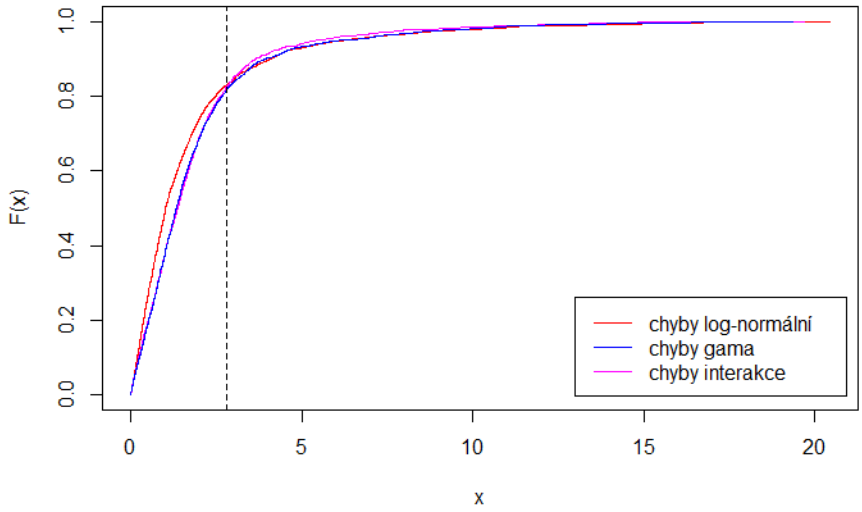


Obrázek 4.1: Boxploty chyb cross-validace pro jednotlivé použité modely, vlevo celé, vpravo vykreslené jen pro chyby menší než 5

Na základě jednotlivých charakteristik polohy bychom zřejmě vybrali model předpokládající log-normální rozdělení pravděpodobnosti, který není nejlepší jen podle extrémních hodnot. Vezmeme-li v potaz také rozptyl, který bychom chtěli co nejmenší, tak jistí už si být nemůžeme, neboť právě v této statistice log-normální rozdělení nejvíce zaostává.

Zajímavý pohled na data nám může dát také vykreslení empirických distribučních funkcí chyb predikce zjištěných pomocí cross-validace, které vidíme na obrázku 4.2. Z něj je patrné, že pro všechny tři modely pravděpodobnost, že chyba bude menší než 2.8, je přibližně rovna 0.85 (hodnota chyby rovná 2.8 je v grafu vyznačená přerušovanou čárou). Tedy naopak, pravděpodobnost toho, že chyba překročí hodnotu 2.8 je u všech modelů přibližně $1 - 0.85 = 0.15$ Pro hodnoty nižší než je tato chyba, bude vždy pravděpodobnost překročení nejnižší u modelu předpokládajícího log-normální rozdělení. Model uvažující interkace věku a PCOS bychom preferovali jen tehdy, kdybychom chtěli snížit pravděpodobnost extrémně velké chyby, tj. pravděpodobnost překročení nějaké chyby větší než 2.8 $\mu g/l$. Například, budeme-li se ptát, jaká je pravděpodobnost, že chyba bude

větší než 5.5 $\mu g/l$, dostaneme v případě modelu s interakcemi mezi PCOS a věkem, že je tato pravděpodobnost přibližně 0.05, zatímco pro zbylé dva modely vychází přibližně 0.06.



Obrázek 4.2: Empirické distribuční funkce chyb predikce zjištěných cross-validací pro jednotlivé modely

Závěr

V úvodní kapitole diplomové práce jsme se stručně seznámili s analyzovanými daty a objasnili si některé klíčové pojmy z medicíny, jako je právě zkoumaný antimülleriánský hormon. Zjistili jsme, proč je tento hormon pro nás důležitý a také jsme vybrali a vhodně transformovali regresory, u nichž jsme se domnívali, že by mohly s hladinou hormonu nějakým způsobem souviset.

Ve druhé části jsme se zaměřili na jednodušší lineární regresní model, který předpokládá, že se závisle proměnná řídí normálním rozdělením. AMH se tímto rozdělením zcela jistě neřídí, nicméně zlogaritmováním naměřených hodnot jsme dostali rozdělení pravděpodobnosti, které už se normálnímu rozdělení podobalo více a pracovali jsme tedy s logaritmovanými hodnotami. Poté jsme se podívali i na další předpoklady lineárního regresního modelu a zjistili, že není splněn předpoklad konstantního rozptylu pro všechna pozorování. Namísto odhadu regresních koeficientů metodou nejmenších čtverců jsme tedy museli pracovat s váženou metodou nejmenších čtverců, pro níž jsme si ukázali také, jak můžeme odhadnout jednotlivé regresní koeficienty. Následně jsme se pokusili najít vhodnou závislost mezi logaritmem AMH a věkem ženy, který jsme považovali za jeden z nejdůležitějších regresorů. K tomu nám posloužilo grafické srovnání klouzavých průměrů hladiny AMH, resp. jejích logaritmovaných hodnot a vyrovnaných hodnot modelu, kde věk byl jedinou vysvětlující proměnnou. Tímto způsobem jsme dospěli k závěru, že nejvhodnější je po částech lineární trend nebo kvadratický trend. Na základě střední čtvercové chyby jsme pak rozhodli ve prospěch po částech lineárního trendu. U ostatních regresorů už jsme uvažovali jen lineární závislost logaritmu hladiny AMH na těchto faktorech a s pomocí Akaikeho in-

formačního kritéria jsme vybrali nejlepší model. Tento model uvažoval kromě věku jako významné regresory také délku menstruačního cyklu, jestli je hladina testosteronu vyšší než 1.5 $nmol/l$, zda žena trpí syndromem PCOS a rok, ve kterém jí byla hladina AMH měřena. Jediný věk má na základě našeho modelu negativní vliv na hladinu AMH v ženském těle. Ženy s nadprůměrně dlouhou délkou menstruačního cyklu mají v průměru vyšší hladinu AMH, stejně jako ženy s PCOS a ženy s hladinou testosteronu vyšší než 1.5 $nmol/l$. V průběhu času navíc hladina testosteronu v populaci stoupá.

Velký pozitivní vliv syndromu PCOS nás vedl k myšlence modelovat hladinu AMH v závislosti na věku zvlášť pro ženy, které tímto syndromem trpí, a pro ženy, které jím netrpí. Tento model však vyžaduje upustit od předpokladu normality závisle proměnné, neboť pracujeme opět s původními koncentracemi AMH v ženském těle. Výsledkem bude, že do modelu přibude jako významný regresor také amenorea (druhá z chorob, které jsme uvažovali jako možné regresory), jejíž přítomnost hladinu hormonu v krvi snižuje.

Ve třetí kapitole jsme si představili, jak vypadá gama rozdělení pravděpodobnosti, a jak můžeme do zobecněného lineárního modelu zahrnout předpoklad, že se závisle proměnná řídí právě tímto rozdělením pravděpodobnosti. Uvedli jsme si vzorce pro výpočet důležitých charakteristik, které potřebujeme k sestavení vhodného modelu, i způsob, jak získat odhady regresních koeficientů. Za splnění určitých přísných předpokladů se nám povedlo tyto odhady odvodit také analyticky, ačkoliv jen pro jednoduchý model s jediným dichotomickým regresorem. V závěru kapitoly jsme pak podobně jako v případě lineárního regresního modelu našli a odhadli nejlepší model podle Akaikeho informačního kritéria. V tomto případě nám jako významné regresory vyšly věk, pro který jsme tentokrát předpokládali kvadratickou závislost, PCOS, délka menstruačního cyklu a rok odběru. Podle modelu působí na AMH přítomnost PCOS, nadprůměrně dlouhý menstruační cyklus i rok odběru opět kladně. Do věku přibližně 26 let předpokládáme také pozitivní vliv věku, teprve poté přijde zlom a hladina hormonu v krvi začíná klesat.

V závěrečné kapitole jsme se podívali na srovnání tří výše uvedených modelů, z nichž každý vynikal v jiném aspektu. Model předpokládající gama rozdělení měl nejnižší největší chybu; model předpokládající log-normální rozdělení má nejlepší výsledky, budeme-li se rozhodovat na základě některého z odhadů polohy rozdělení chyb: první kvartil, medián, třetí kvartil či průměr; a model s interakcemi mezi PCOS a věkem má nejmenší pravděpodobnost chyby větší než 2.8 $\mu g/l$ (pomineme-li extrémní chyby, v nichž, jak jsme uvedli, byl nejlepší model předpokládající gama rozdělení závisle proměnné).

Literatura

[1] AMH | Sanatorium Helios CZ, https://www.sanatoriumhelios.cz/amh/ [cit. 4. 1. 2019].

[2] Endocare – Syndrom polycystických ovárií, http://endokrinologie-obezitologie.cz/cs/clanky/poruchy-menstruace/syndrom-polycystickych-ovarii/ [cit. 4. 1. 2019].

[3] amenorea | Velký lékařský slovník On-Line, http://lekarske.slovniky.cz/lexikon-pojem/amenorea-amenorrhoea-4 [cit. 4. 1. 2019].

[4] Hron, K., Kunderová, P.: *Základy počtu pravděpodobnosti a metod matematické statistiky* (1.vydání). Univerzita Palackého v Olomouci, Olomouc, 2015. ISBN: 978-80-244-3396-7

[5] Fišerová, E..: *Lineární statistické modely* (1. vydání). Univerzita Palackého v Olomouci, Olomouc, 2013. ISBN: 978-80-244-3402-5

[6] McCullagh, P., Nelder, J. A.: *Generalized Linear Models* (2nd edition), New York: Chapman and Hall, 1989. ISBN 9780412317606

[7] Generalized linear models. Statistics and Probability - MSU - Department of Statistics and Probability [online]. [cit. 2019-04-01]. Dostupné z: https://www.stt.msu.edu/users/pszhong/Lecture_12_Spring_2017.pdf