MA012 Statistika II

cvičení 10–11

Ondřej Pokora (pokora@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

(podzim 2015)

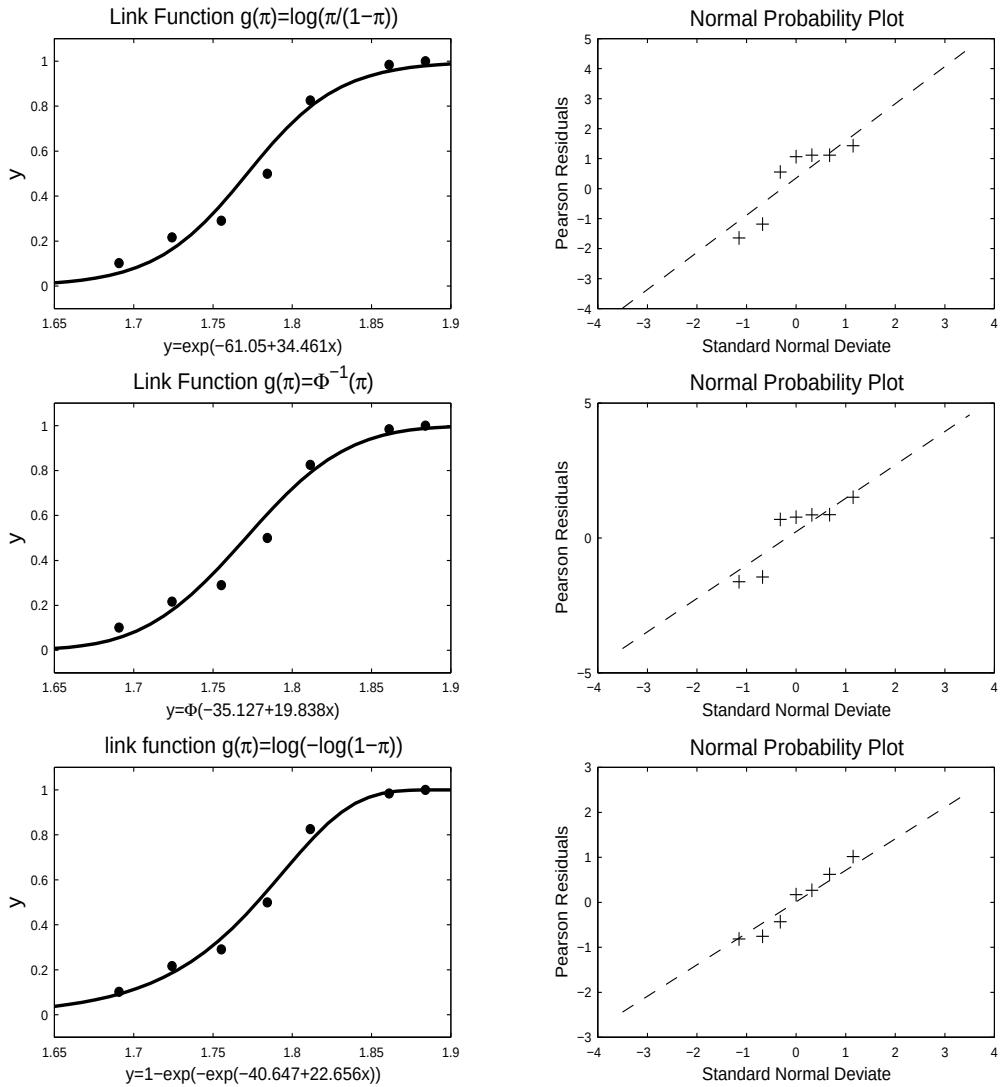
Příklad 1

V souboru `beetle.csv` *jsou uvedeny údaje o úmrtnosti Potemníka skladištního (Tribolium confusum) v reakci na sirouhlík CS_2. Datový soubor obsahuje tyto proměnné*

`dose`	*množství sirouhlíku* (mg/l)
`population`	*počet kusů ve zkoumaném vzorku*
`killed`	*počet mrtvých kusů ve zkoumaném vzorku*

Modelujte závislost úmrtnosti na množství CS_2.

Řešení. Pro modelování závislosti použijeme logistický model, probitový model a model s komplementární log-log linkovací funkcí.



Obrázek: Modely pro úmrtnost *Potemníka skladištního*.

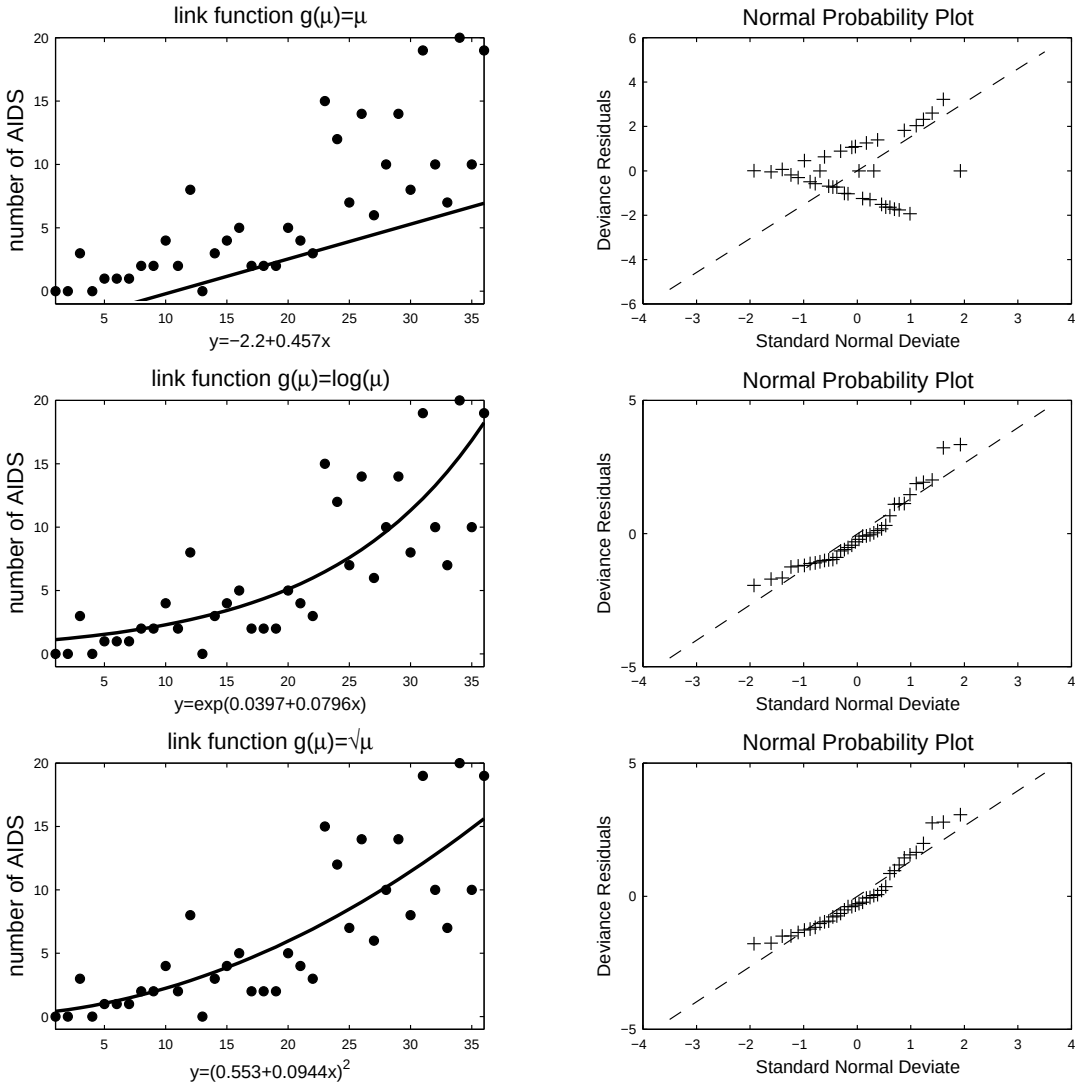
Příklad 2

V souboru `aids.csv` *jsou uvedeny údaje o počtech nových případů AIDS ve Velké Británii za období prosinec 1982 až listopad 1985. Datový soubor obsahuje tyto proměnné*

month	*měsíc*
year	*rok*
number	*počet nových případů AIDS*

Modelujte závislost počtu nových případů AIDS na čase.

Řešení. Pro modelování závislosti použijeme lineární model, log-lineární model a odmocninový model.



Obrázek: Modely pro výskyt nových onemocnění AIDS ve Velké Británii.

Příklad 3

V souboru `sharks.csv` *jsou k dispozici data, která popisují počty napadení žraloky na Floridě v letech 1946 až 1999. Známe také velikost populace. Datový soubor obsahuje tyto proměnné:*

`Year`	*rok*
`Population`	*velikost populace*
`Attacks`	*počet napadení žraloky*
`Fatalities`	*počet úmrtí způsobených žraloky*

Nejprve vykreslete bodový graf počtu napadení na 1 milión obyvatel v závislosti na čase. Pro modelování použijte binomický i poissonovský model s kanonickou linkovací funkcí. Pro matici plánu uvažujte kubický polynom v proměnné `Year`.

Příklad 3

Predikce obou modelů i s intervalem spolehlivosti pro regresní funkci vykreslete do obrázku. Zkoumejte také, jestli nenastal problém příliš velkého nebo příliš malého rozptylu. Pokud ano, předefinujte model a výsledky znovu vykreslete do obrázku. Pomocí výsledného modelu odhadněte, kolik útoků (na 1 milión obyvatel) způsobí žraloci na Floridě v roce 2013 a také v jakém intervalu se tato hodnota s 95% pravděpodobností bude pohybovat.

[Nastal problém příliš velkého rozptylu. Odhad: 33,96 útoků na 1 milión obyvatel, interval spolehlivosti: $[3,207; 359,55]$.]

Příklad 4

V souboru `car_income.csv` *jsou uvedeny údaje o koupi nového auta během posledních 12-ti měsíců v závislosti na příjmu domácnosti a stáří původního auta. Datový soubor obsahuje tyto proměnné:*

`purchase`	*indikátor nákupu nového auta (1 – ano, 0 – ne)*
`income`	*roční příjem domácnosti (v tis. dolarů)*
`age`	*stáří původního auta (roky)*

Nejprve vykreslete závislosti proměnné `purchase` *na ostatních. Pro modelování závislosti nalezněte vhodný logistický model. Jsou všechny proměnné statisticky významné? Znovu modelujte s použitím proměnné* `age` *jako* `factor`*. Opět sledujte statistickou významnost* `age`*. Vyzkoušejte tuto proměnnou zakomponovat do modelu jako* `factor` *s méně úrovněmi. Výsledky vykreslete do obrázku.*

Příklad 5

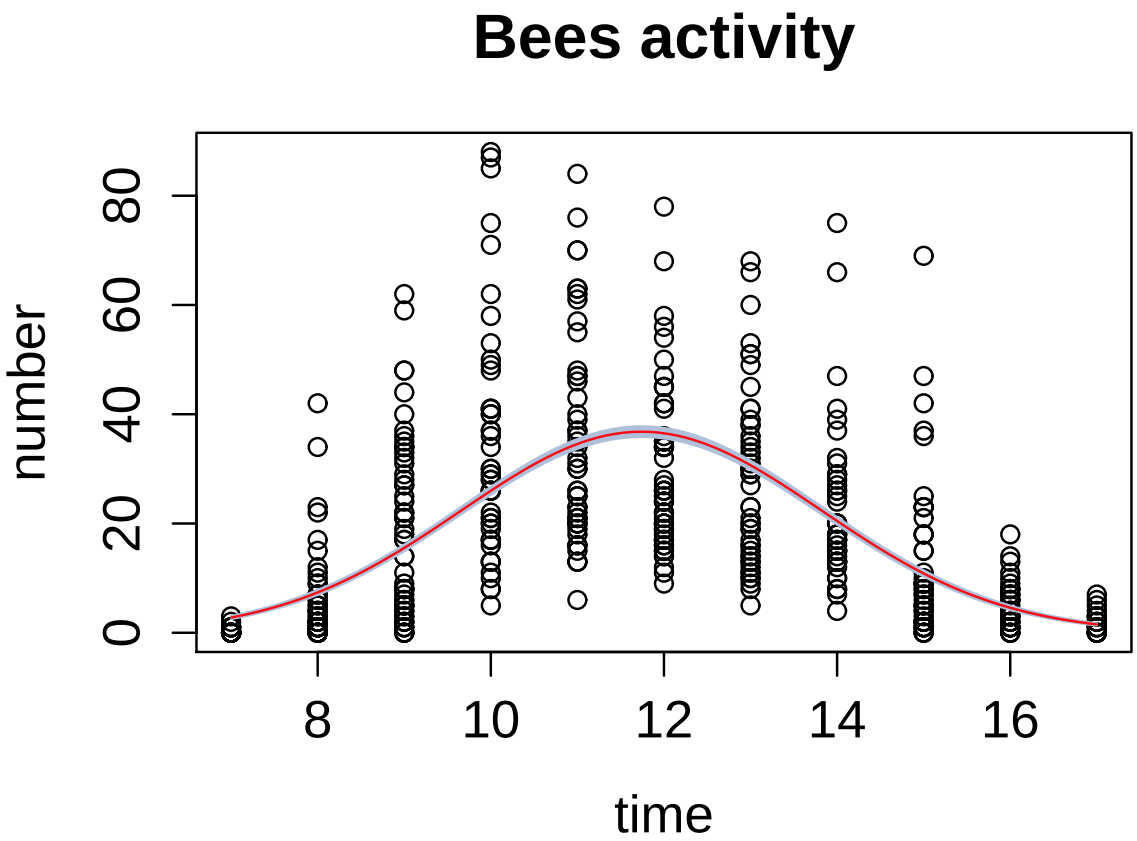
V souboru `bees.csv` *jsou uvedeny údaje o aktivitě včel v závislosti na čase. Jednou z důležitých charakteristik při zkoumání včelí aktivity je počet včel, které opustí úl kvůli práci ve vnějším prostředí. Studie se zabývala měřením této veličiny během několika slunečných dní v závislosti na čase během dne. Datový soubor obsahuje tyto proměnné*

`number`	*počet včel, které opustily úl*
`time`	*čas, kdy byl tento údaj zaznamenán*

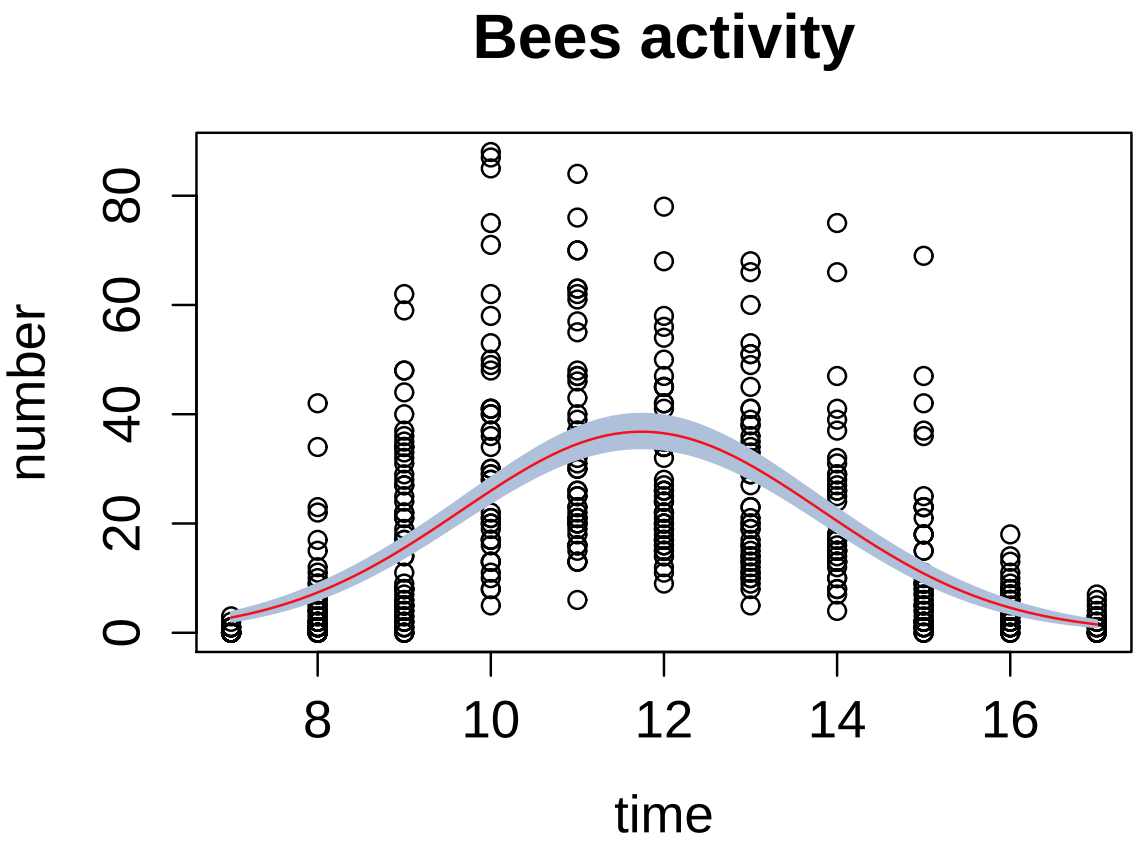
Modelujte závislost počtu včel, které opustí úl, na čase během dne.

řešení

Pro modelování závislosti použijeme poissonovský model s kanonickou linkovací funkcí. Do modelu vstupuje jediná vysvětlující proměnná `time` a přidáme také její druhou mocninu.
Hodnota reziduální deviance (4 879,3) je nepoměrně vyšší než počet stupňů volnosti (501). Je zřejmé, že došlo k „overdispersion“ a v jazyce R je třeba volit `family=quasipoisson`. Použití této volby neovlivňuje odhady koeficientů, ale mění jejich odhady variability, což se projeví např. v intervalu spolehlivosti.



Obrázek: Odhad regresní funkce **bez** vyrovnání se s problematikou velkého rozptylu.



Obrázek: Odhad regresní funkce **s** vyrovnáním se s problematikou velkého rozptylu.

Příklad 6

V souboru `heart.csv` *jsou uvedena data o přítomnosti infarktu myokardu v závislosti na věku pacienta. Datový soubor obsahuje tyto proměnné:*

`age`	*věk pacienta (roky)*
`chd`	*indikátor infarktu (1 – nastal, 0 – nenastal)*

Pro modelování závislosti použijte logistický model, probitový model a model s komplementární log-log linkovací funkcí. Výsledky vykreslete do obrázku.

Příklad 7

V souboru `nemocnice.csv` *jsou uvedeny údaje o zotavení pacientů v závislosti na závažnosti onemocnění a nemocnici, ve které se léčili. Datový soubor obsahuje tyto proměnné:*

`Infection_Severity`	*vážnost onemocnění*
`Treatment_Outcome`	*indikátor uzdravení (1 – zdravý, 0 – smrt)*
`Hospital`	*typ nemocnice (1, 2, 3)*

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku.

Příklad 8

V souboru `cancer.csv` *jsou uvedeny údaje o počtu onemocnění rakovinou kůže u žen v závislosti na věku a oblasti v USA, ve které pacientky žily. Datový soubor obsahuje tyto proměnné:*

`Cases`	*počet onemocnění*
`Town`	*město (0 – Minneapolis (Minnesota), 1 – Dallas (Texas))*
`Age`	*věková skupina pacientky*
`Population`	*celkový počet žen dané věkové skupiny v příslušném městě*

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku. Porovnejte pravděpodobnost vzniku onemocnění u 60-ti leté pacientky žijící v Minneapolisu s pravděpodobností pro stejně starou pacientku žijící v Dallasu.

[Minneapolis: 0.00117, Dallas: 0.00276.]

Příklad 9

V souboru `druhy.csv` *jsou k dispozici data, která se týkají dlouhodobého zemědělského experimentu. Bylo sledováno 90 pozemků (pastvin) o rozloze 25 m* $\times$ *25 m, lišících se v biomase, pH půdy a druhové bohatosti (počet rostlinných druhů na celém pozemku). Je dobře známo, že s rostoucí biomasou dochází k poklesu druhové bohatosti. Ale zůstává otázka, zda rychlost poklesu nesouvisí s úrovní pH v půdě. Proto byly jednotlivé pozemky klasifikovány podle hodnoty pH v půdě do tří úrovní (nízká, střední a vysoká úroveň) a do experimentu bylo vybráno vždy po 30 pozemcích pro každou úroveň. Spojitá veličina* `Biomass` *je dlouhodobým průměrem naměřených červnových hodnot biomasy. Datový soubor obsahuje tyto proměnné:*

`pH`	*úroveň pH v půdě (*`low` *– nízká,* `mid` *– střední,* `high` *– vysoká)*
`Biomass`	*množství biomasy*
`species`	*počet rostlinných druhů*

Příklad 9

Nejprve vykreslete závislosti proměnné `species` *na ostatních. Pro modelování závislosti nalezněte vhodný poissonovský model. Vyzkoušejte postupně logaritmickou, identickou a odmocninovou linkovací funkci. Jsou všechny proměnné statisticky významné? Pokud ne, zkuste modely zjednodušit a pomocí analýzy deviance rozhodněte, zda takové zjednodušení je možné. Získané výsledné modely vykreslete do obrázku. Pomocí všech modelů odhadněte počet rostlinných druhů na pozemku s hodnotou biomasy 9 a střední úrovní pH v půdě.*

[Odhady počtu druhů pro log link: 8,895, identity link: 4,513, sqrt link: 7,414.]

Příklad 10

V souboru `motak.csv` *jsou uložena data o lovu tetřeva dravcem jménem Moták pilich (Circus cyaneus) v závislosti na výskytu tetřeva. Označme* Y_i *procento zkonzumovaných tetřevů a* x_i *počet tetřevů v dané oblasti. Teorie zabývající se chováním těchto dravců navrhují k modelování použít vztahu*

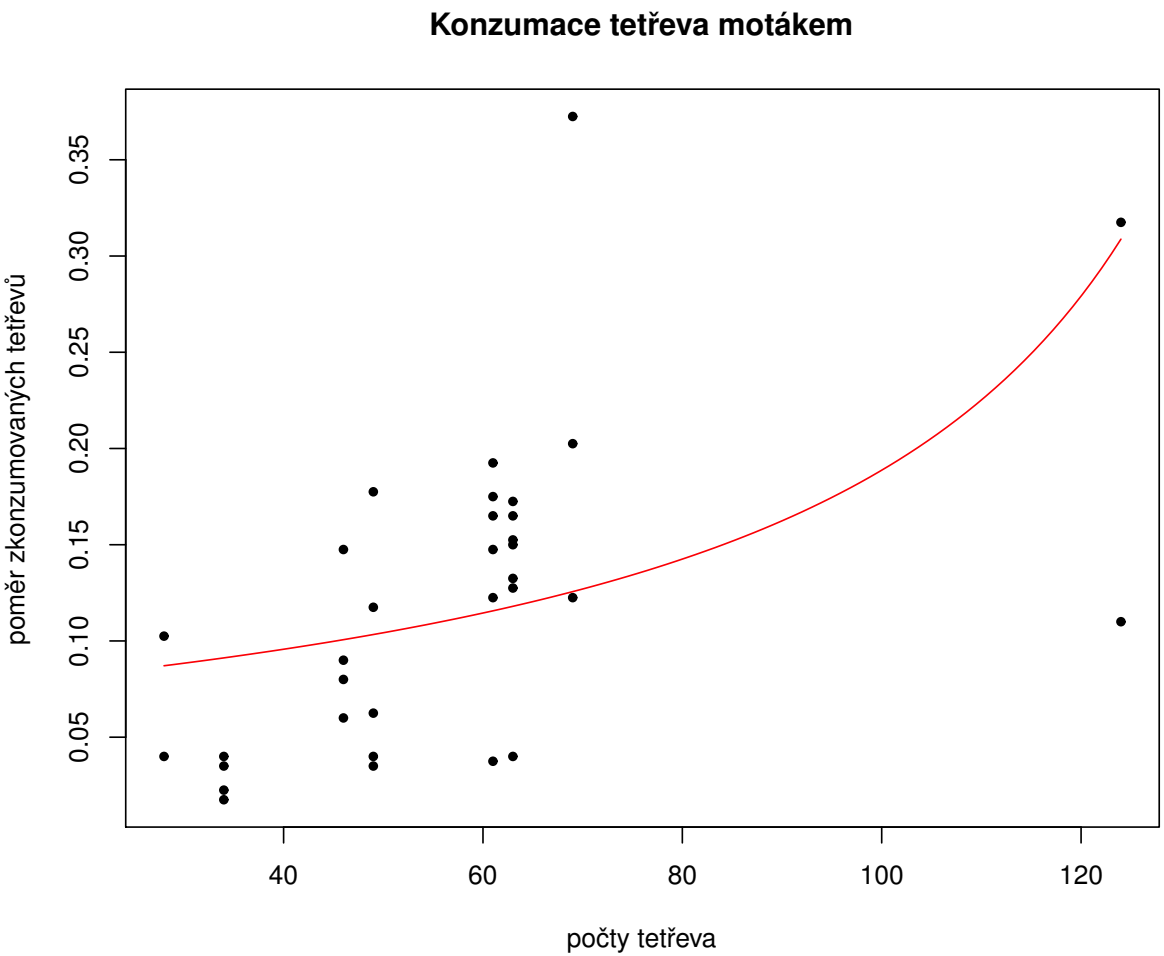
$$E(Y_i) = \mu_i = \frac{\alpha x_i^3}{\delta + x_i^3},$$

kde Y_i *má Gamma rozdělení. Je tedy třeba odhadnout neznámé parametry* α *a* δ*. Užitím linkovací funkce* `inverse` *dostáváme*

$$\frac{1}{\mu_i} = \frac{1}{\alpha} + \frac{\delta}{\alpha x_i^3}.$$

Definování nových parametrů $\beta_0 = 1/\alpha$ *a* $\beta_1 = \delta/\alpha$ *dostáváme lineární vztah*

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 \frac{1}{x_i^3}.$$



Obrázek: Aplikace Gamma regrese s linkovací funkcí $g(\mu) = 1/\mu$ na data `motak`.