M7222 – 4. cvičení : **GLM04a** (*Problémy s příliš malou či příliš velkou variabilitou:* underdispersion, overdispersion)

Mějme náhodný výběr $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ z rozdělení exponenciálního typu, který se řídí *GLM* modelem, tj. se sdruženou hustotou pravděpodobnosti či sdruženou pravděpodobnostní funkcí tvaru

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \theta_i) = \exp\left\{ \sum_{i=1}^{n} \left[\frac{y_i\theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right] \right\}$$

tj. jde o regulární hustotu exponenciálního typu v kanonické škálové formě. Předpokládejme, že pro hustotu exponenciálního typu platí:

$$\psi_i(\phi) = \frac{\phi}{\omega_i},$$

kde $\omega_i > 0$ jsou **známé apriorní váhy** a $\phi > 0$ je **neznámý rušivý parametr**, který se též nazývá **škálovým** či **rozptylovým** parametrem.

Pak při testování vhodnosti modelu hraje velmi důležitou roli tzv. (škálová) deviace, kterou můžeme vyjádřit takto

$$\begin{aligned} D &= 2\left[l^*(\widehat{\boldsymbol{\beta}}_{max}; \mathbf{Y}) - l^*(\widehat{\boldsymbol{\beta}}; \mathbf{Y}) \right] \\ &= 2\sum_{i=1}^{n} \left\{ \frac{\omega_i \left[Y_i\hat{\theta}_{i,max} - \gamma(\hat{\theta}_{i,max}) \right]}{\phi} + d(Y_i, \phi) - \frac{\omega_i \left[Y_i\hat{\theta}_i - \gamma(\hat{\theta}_i) \right]}{\phi} - d(Y_i, \phi) \right\} \\ &= \frac{1}{\phi} 2\sum_{i=1}^{n} \omega_i \left[Y_i(\hat{\theta}_{i,max} - \hat{\theta}_i) - \gamma(\hat{\theta}_{i,max}) + \gamma(\hat{\theta}_i) \right] \\ &= \frac{1}{\phi} D^* \end{aligned}$$

a D^* nazveme **neškálovanou deviací** (unscaled deviance). Protože platí

$$D = \frac{1}{\phi} D^* \quad \overset{A}{\sim} \quad \chi^2(n-m) \qquad \Rightarrow \qquad ED = \frac{1}{\phi} ED^* \approx n - m,$$

neboť střední hodnota u χ^2 je rovna počtu stupňů volnosti, pak

$$\hat{\phi}_{D^*} = \frac{D^*}{n-m} \, .$$

Další často používanou mírou vhodnosti modelu je tzv. **zobecněná Pearsonova statistika**

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \quad \overset{A}{\sim} \quad \chi^2(n-m)$$

a proto dalším momentovým odhadem založeným na této statistice je

$$\hat{\phi}_{X^2} = \frac{X^2}{n-m} \, .$$

Přehled rozptylových parametrů a neškálovaných deviací pro rozdělení exponenciálního typu je dán v následující tabulce.

Rozdělení	ϕ	D^*
Normální rozdělení	σ^2	$\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2$
Poissonovo rozdělení	1	$2\sum_{i=1}^{n}\left[Y_i ln\frac{Y_i}{\hat{\mu}_i}-(Y_i-\hat{\mu}_i)\right]$
Binomické rozdělení	1	$2\sum_{i=1}^{n}\left[Y_i\ln\frac{y_i}{\hat{\mu}_i}+(n_i-Y_i)\ln\frac{n_i-Y_i}{n_i-\hat{\mu}_i}\right]$
Gamma rozdělení	$\frac{1}{\alpha}$	$2\sum_{i=1}^{n}\left[\frac{Y_i-\hat{\mu}_i}{\hat{\mu}_i}-ln\frac{Y_i}{\hat{\mu}_i}\right]$

Problém s příliš velkou či malou variabilitou se týká těch rozdělení, u kterých má být scale parametr roven jedné, tj. binomického a Poissonova rozdělení. Proto si nejprve tato dvě rozdělení připomeňme.

Náhodná veličina s binomickým rozdělením $Z = nY \sim Bi(n,\pi)$, kde $n \in \mathbb{N}, \pi \in (0,1)$, má

$$f_Z(z) = \binom{n}{z}\pi^z(1-\pi)^{n-z} = \exp\left\{z\ln\left(\frac{\pi}{1-\pi}\right)+n\ln(1-\pi)+\ln\binom{n}{z}\right\} \quad \text{pro} \quad z=0,\dots,n,$$

přičemž

$$EZ = \mu = n\pi \qquad \text{a} \qquad DZ = n\pi(1-\pi).$$

Tedy ***přirozený parametr*** $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$

rozptylová funkce $V(\mu) = \mu(1-\mu)$

scale factor $\phi = 1$

váhy $\omega = n.$

Vidíme, že platí

$$DZ < EZ \qquad \text{nebot'} \qquad DZ = \underbrace{n\pi}_{EZ}(1-\pi) \quad \text{a} \quad (1-\pi)\in(0,1).$$

Naproti tomu náhodná veličina s Poissonovým rozdělením $Y \sim Po(\lambda)$, kde $\lambda > 0$ má

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp\{y\ln\lambda-\lambda-\ln y!\} \quad \text{pro} \quad y=0,1,2,\dots$$

přičemž

$$EY = \mu = \lambda \qquad \text{a} \qquad DY = \lambda.$$

Tedy ***přirozený parametr*** $\theta = \ln\lambda$

rozptylová funkce $V(\mu) = \mu$

scale factor $\phi = 1$

váhy $\omega = 1.$

Pro Poissonovo rozdělení tedy platí, že

$$DY = EY.$$

Pokud pro reálná data dojde k tomu, že například pro binomické či Poissonovo rozdělení je rozptyl větší než střední hodnota, pak jde o *overdispersion*. Pokud je například u dat, pro která jsme předpokládali Poissonovo rozdělení, rozptyl naopak menší než střední hodnota, pak jde o *underdispersion*.

V těchto případech není hodnota disperzního (*scale*) parametru ϕ (jakožto poměru $\frac{DY}{EY}$) rovna 1.

Ve výpisu výsledků modelu nás na tuto situaci upozorní výrazně větší (menší) hodnota reziduální (tedy nevysvětlené) deviace ve srovnání s reziduálním počtem stupňů volnosti, což je střední hodnota χ^2 rozdělení.

Existuje řada možných vysvětlení, proč k tomu došlo. Tak například v biologických studiích může být *overdispersion* důsledkem agregovaného výskytu organismů. Nebo je tento jev důsledkem závislosti v datech, které standardní model nepředpokládá. Může se také stát že přirozený parametr není stálý, ale mění se náhodně mezi jedinci. Příliš malý či velký rozptyl může vzniknout také nezařazením některé důležité vysvětlující proměnné.

V prostředí R je k řešení tohoto problému k dispozici modifikovaná volba pro třídu exponenciálního rozdělení. V případě binomického rozdělení máme možnost volby

`family=quasibinomial`

a pro Poissonovo rozdělení

`family=quasipoisson`.

Nejde o nový typ exponenciálního rozdělení, ale o změnu ve výpočtu druhého momentu, pro jehož odhad se použije jednoduchý momentový odhad disperzního parametru ϕ.

Výsledná korekce rozptylu je pak důležitá při testování hypotéz, neboť zohledňuje vyšší/nižší variabilitu v datech a zabraňuje tak nadbytku/nedostatku falešně pozitivních výsledků testů hypotéz o parametrech modelu.

Hodnota parametru ϕ se stanoví z poměru Pearsonových reziduí a reziduálního počtu stupňů volnosti, tj. pomocí $\hat{\phi}_{X^2}$.

Vrátíme se k příkladu *The Working Activities of Bees*. Nejprve načteme popis dat, následně samotná data, nakonec data vykreslíme.

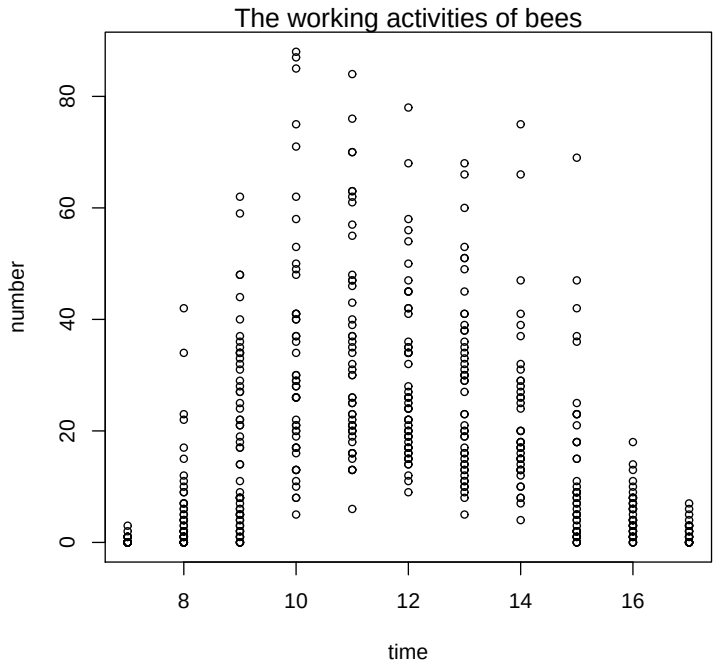
```
> fileTxt <- paste(data.library, "bees.txt", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))
```

```
 [1] "The working activities of bees"
 [2] "=============================="
 [3] "http://www.math.tau.ac.il/~felix/GENLIN/Bees.dat "
 [4] ""
 [5] "The file contains the data about \"working activities\" "
 [6] "of bees in the bee-hive (Hebrew: \"kaveret\") as a function "
 [7] "of time of the day. One of the important characteristics "
 [8] "of \"working activities\" is the number of bees leaving "
 [9] "the bee-hive for outside activities. "
[10] "The data collected during several successive non-rainy days "
[11] "contain the number of bees that left the bee-hive and the time "
[12] "of the day (in hours). "
```

```
> close(con)
> fileDat <- paste(data.library, "bees.dat", sep = "")
> data <- read.table(fileDat, header = FALSE)
> names(data) <- c("number", "time")
> str(data)
```

```
,data.frame,:   504 obs. of  2 variables:
 $ number: int  34 13 11 32 39 36 34 20 16 35 ...
 $ time  : int  9 10 12 13 14 15 9 10 12 13 ...
```

```
> with(data, plot(number ~ time, cex = 0.75))
> mtext(popis[1], cex = 1.25)
```



Obrázek 1: Bodový graf dat *The Working Activities of Bees*.

Protože závisle proměnná `number` značí počty včel, budeme předpokládat, že její rozdělení je Poissonovo. Jako linkovací funkci zvolíme kanonickou, tj. logaritmus. Pomocí GLM modelu budeme zkoumat vztah mezi proměnnými `number` a `time`.

```
> m1 <- glm(number ~ time + I(time^2), data = data, family = poisson)
> summary(m1)
```

```
Call:
glm(formula = number ~ time + I(time^2), family = poisson, data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.009  -2.691  -1.232   1.317  11.789
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.235733   0.285242  -42.90   <2e-16 ***
time          2.698642   0.049285   54.76   <2e-16 ***
I(time^2)    -0.114931   0.002096  -54.84   <2e-16 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9305.5  on 503  degrees of freedom
Residual deviance: 4879.3  on 501  degrees of freedom
AIC: 6830.6

Number of Fisher Scoring iterations: 6
```

Všimněme si, že hodnota reziduální deviace je nepoměrně vyšší než počet stupňů volnosti, což je střední hodnota χ^2 rozdělení.

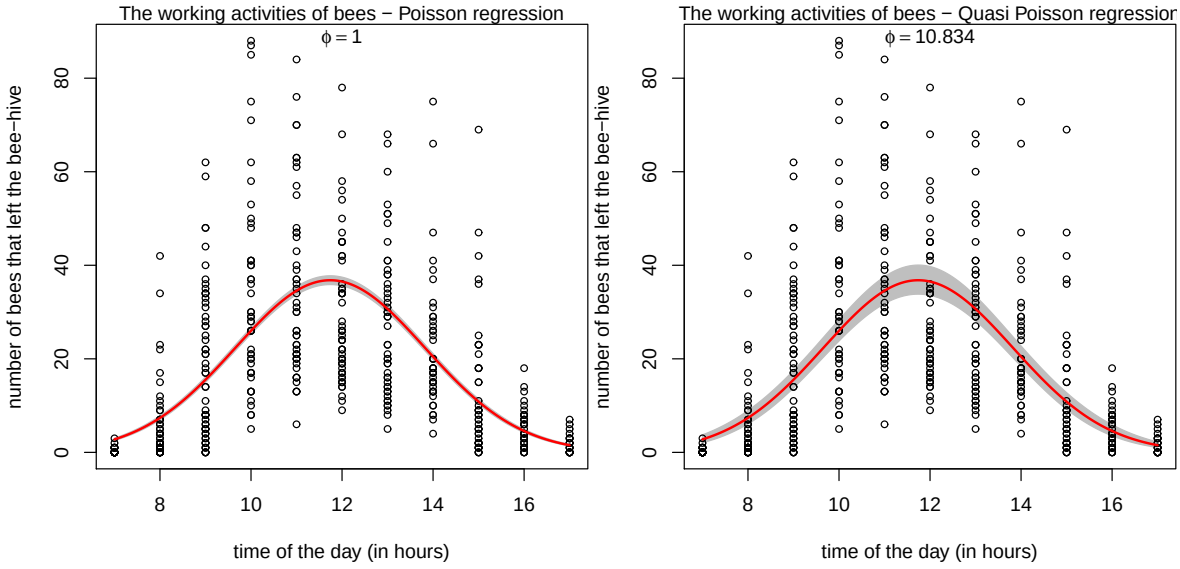
```
> (Scale <- m1$deviance/m1$df.residual)

[1] 9.739058
```

Vidíme,že tato hodnota je téměř desetkrát vyšší. Zvolíme tedy variantu s volbou `family=quasipoisson`.

```
> m1q <- glm(number ~ time + I(time^2), data = data, family = quasipoisson)
> summary(m1q)


Call:
glm(formula = number ~ time + I(time^2), family = quasipoisson,
    data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.009  -2.691  -1.232   1.317  11.789

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.235733   0.938884  -13.03   <2e-16 ***
time          2.698642   0.162223   16.64   <2e-16 ***
I(time^2)    -0.114931   0.006898  -16.66   <2e-16 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

(Dispersion parameter for quasipoisson family taken to be 10.83416)

    Null deviance: 9305.5  on 503  degrees of freedom
Residual deviance: 4879.3  on 501  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Vidíme, že použití volby `family=quasipoisson` neovlivňuje odhady koeficientů, ale mění jejich odhady variability. Na závěr tohoto příkladu ještě graficky srovnáme oba výsledky.

```
> par(mfrow = c(1, 2), mar = c(4, 4, 2, 0) + 0.05)
> ab <- range(data$time)
> xx <- seq(ab[1], ab[2], length.out = 200)
> x <- c(xx, rev(xx))
> yy <- predict(m1, list(time = xx), type = "response")
> predicted.log <- predict(m1, list(time = xx), type = "link", se = T)
> CI.L.log <- exp(predicted.log$fit - 1.96 * predicted.log$se.fit)
> CI.H.log <- exp(predicted.log$fit + 1.96 * predicted.log$se.fit)
> y <- c(CI.L.log, rev(CI.H.log))
> plot(data$time, data$number, type = "n", ylab = "number of bees that left the bee-hive",
    xlab = "time of the day (in hours)")
> polygon(x, y, col = "gray75", border = "gray75")
> points(data$time, data$number, cex = 0.75)
> lines(xx, yy, col = "red", lwd = 2)
> mtext(paste(popis[1], "-", "Poisson regression"), cex = 1.025)
> phi <- round(summary(m1)$dispersion, 3)
> mtext(text = bquote(phi == .(phi)), side = 3, line = -1, cex = 1)
> yy <- predict(m1q, list(time = xx), type = "response")
> predicted.log <- predict(m1q, list(time = xx), type = "link", se = T)
> CI.L.log <- exp(predicted.log$fit - 1.96 * predicted.log$se.fit)
> CI.H.log <- exp(predicted.log$fit + 1.96 * predicted.log$se.fit)
> y <- c(CI.L.log, rev(CI.H.log))
> plot(data$time, data$number, type = "n", ylab = "number of bees that left the bee-hive",
    xlab = "time of the day (in hours)")
> polygon(x, y, col = "gray75", border = "gray75")
> points(data$time, data$number, cex = 0.75)
> lines(xx, yy, col = "red", lwd = 2)
> mtext(paste(popis[1], "-", "Quasi Poisson regression"), cex = 1.025)
> phi <- round(summary(m1q)$dispersion, 3)
> mtext(text = bquote(phi == .(phi)), side = 3, line = -1, cex = 1)
```



Obrázek 2: Srovnání výsledků bez a s vyrovnáním se s problematikou přílis velkého rozptylu.