M7222 – 3. CVIČENÍ : **GLM03a** (*The Working Activities of Bees*)

Popis dat je v souboru `bees.txt`, samotná data jsou uložena v souboru `bees.dat`. Nejprve načteme popisný soubor pomocí příkazu `readLines()` (kterému musí předcházet příkaz `file()` a po něm následuje příkaz `close()`). Protože je příkaz v závorkách, ihned se zobrazí obsah souboru.

```
> 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)
```

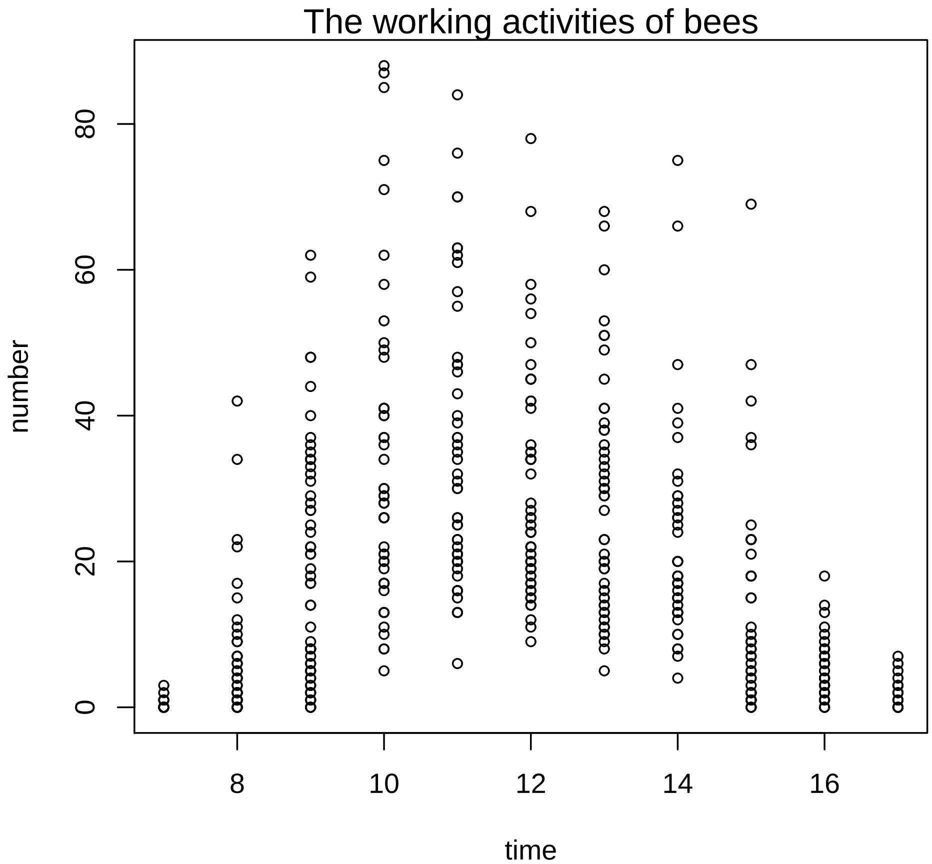
Nyní načteme datový soubor pomocí příkazu `read.table()` a pojmenujeme proměnné. Příkazem `str()` vypíšeme strukturu datového rámce.

```
> 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 ...
```

Abychom získali grafickou představu o datech vykreslíme je.

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

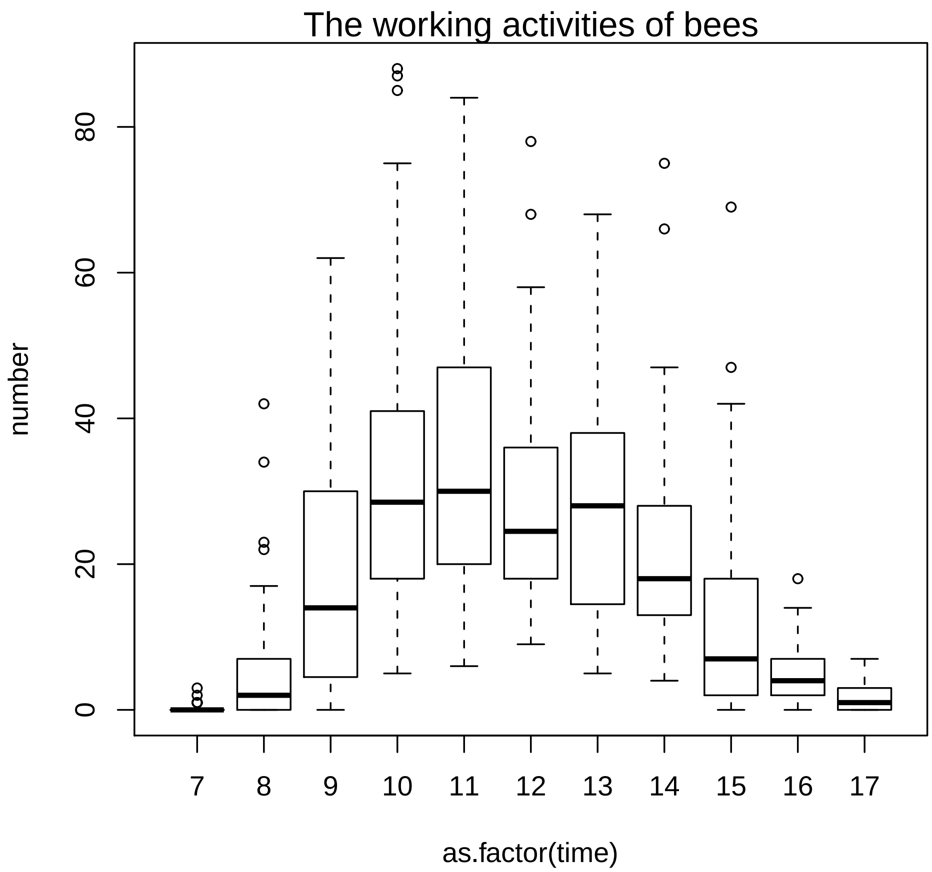


Obrázek 1: Bodový graf vstupních dat.

Z grafu je zřetelně vidět, že variabilita proměnné `number` je funkcí proměnné `time`.

Abychom to lépe demonstrovali vykreslíme následující graf.

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



Obrázek 2: Boxploty pro různé hodnoty proměnné `time`.

Nyní budeme zkoumat vztah mezi proměnnými `number` a `time` pomocí GLM modelu.

Protože závisle proměnná `number` značí počty včel, budeme předpokládat, že její rozdělení je Poissonovo. Jako linkovací funkci zvolíme kanonickou, tj. logaritmus.

```
> 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
```

Vidíme, že jednotlivé kovariáty jsou statisticky významné.

Vhodnost modelu ověříme pomocí Waldova testu.

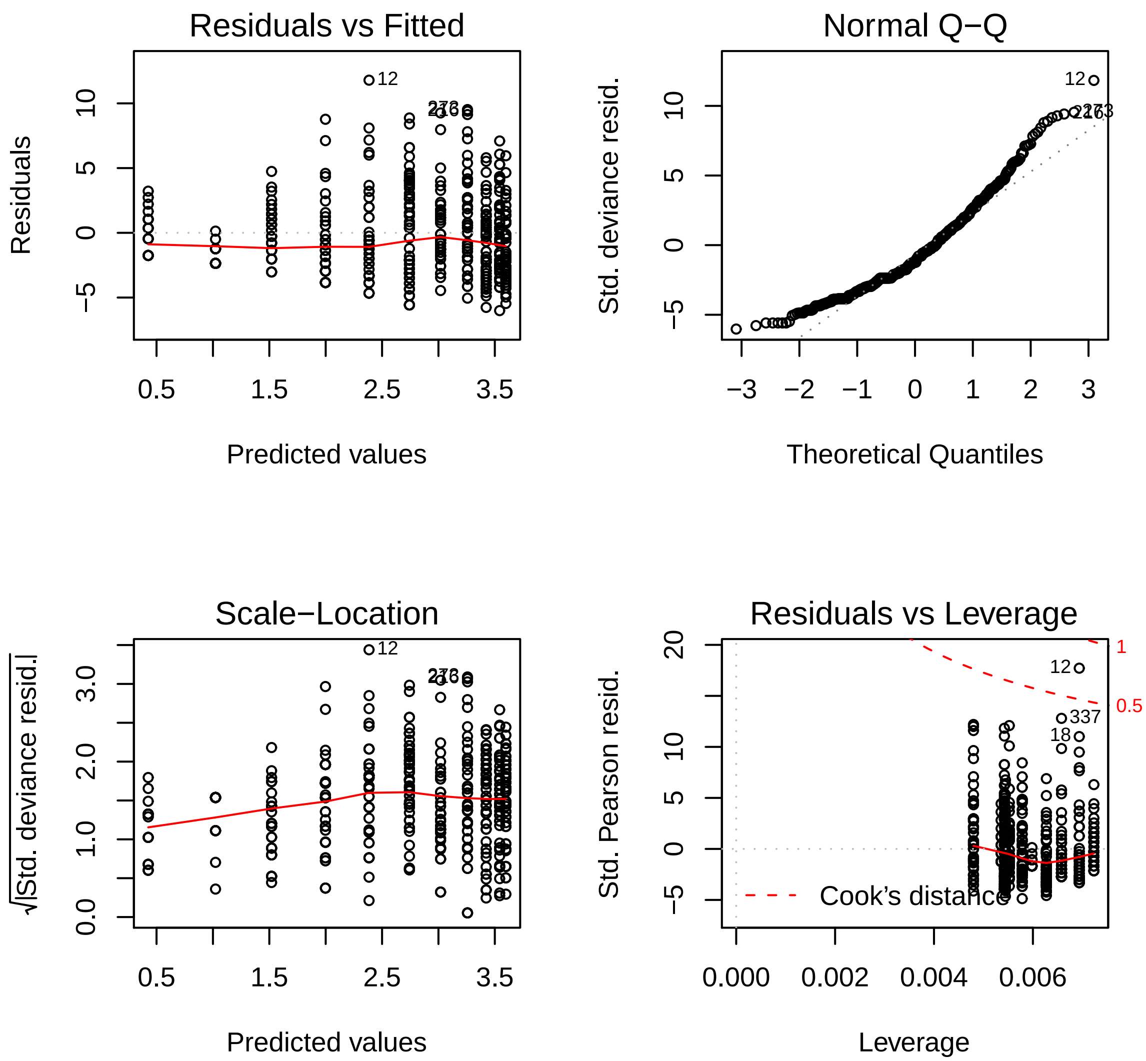
```
> library(lmtest)
> waldtest(m1, test = "Chisq")
```

```
Wald test

Model 1: number ~ time + I(time^2)
Model 2: number ~ 1
  Res.Df Df  Chisq Pr(>Chisq)
1    501
2    503 -2 3012.3  < 2.2e-16 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1
```

Z výsledku je zřejmé, že oproti nulovému modelu se přidáním kovariát `time` a `I(time^2)` model výrazně zlepšil. Nesmíme zapomenout také na anlýzu reziduí.

```
> par(mfrow = c(2, 2))
> plot(m1, cex = 0.75)
```



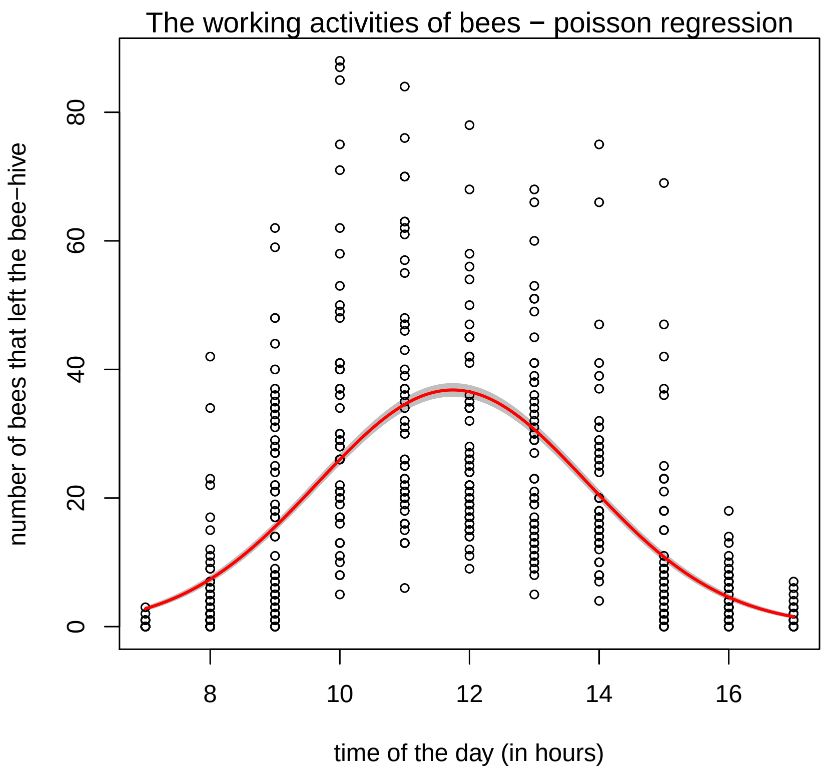
Obrázek 3: Analýza reziduí pro poissonovskou regresi s kvadratickým trendem v lineárním prediktoru pro data *The Working Activities of Bees*.

Z grafů je vidět, že analýza reziduí nedopadla nejlépe. Další kovariáty však nemáme k dispozici.

Na závěr tohoto příkladu ještě graficky znázorníme výsledek. Do grafu také zakreslíme asymptotické intervaly spolehlivosti kolem střední hodnoty.

```
> ab <- range(data$time)
> xx <- seq(ab[1], ab[2], length.out = 200)
> yy <- predict(m1, list(time = xx), type = "response")
> predicted.log <- predict(m1, list(time = xx), type = "link",
     se = T)
> CI.L.log <- exp(predicted.log$fit - 1.96 * predicted.log$se.fit)
> CI.H.log <- exp(predicted.log$fit + 1.96 * predicted.log$se.fit)
> x <- c(xx, rev(xx))
```

```
> 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 = 1.125)
```



Obrázek 4: Poissonovská regrese spolu s asymptotickými intervaly spolehlivosti.