M7222 – 4. cvičení : **GLM04b** (*Vztah mezi Poissonovým a binomických rozdělením*)

Připomeňme, že pomocí Poissonova rozdělení $Po(\lambda)$ lze dobře **aproximovat binomické rozdělení** $Bi(n, \pi)$ za podmínek, že

$$n \to \infty \quad \wedge \quad \pi \to 0 \quad \wedge \quad n\pi = \lambda < \infty \ ,$$

obvykle se doporučuje $n > 30$ a $\pi < 0.1$.

Chceme-li tedy aproximovat binomické rozdělení $Bi(n_i, \pi_i)$ (kde n_i jsou dostatečně velká a π_i dostatečně malá) pomocí rozdělení Poissonova $Y_i \sim Po(\lambda_i = n_i\pi_i)$ a přitom použijeme **logaritmickou linkovací funkci**, platí

$$\lambda_i = n_i\pi_i = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \quad \Rightarrow \quad \log(\lambda_i) = \underbrace{\log(n_i)}_{\text{tzv. offset}} + \log(\pi_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}.$$

Příklad: Shark Attacks

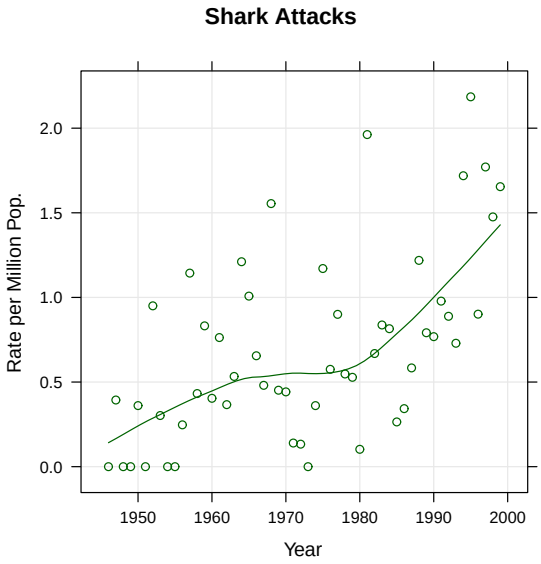
Máme k dispozici data, která popisují počty napadení žraloky na Floridě v letech 1945 až 2000. Vedle toho známe i velikost populace. Počty napadení jsou velmi malé vzhledem k velikosti populace, takže místo binomického rozdělení budeme uvažovat Poissonovo rozdělení.

Data načteme, prohledneme a vykreslíme. Aby grafická informace měla smysl, místo počtu napadení vykreslíme za jednotlivé roky podíly vzhledem k velikosti populace.

```
> df <- read.csv(paste(data.library, "floridashark.csv", sep = ""))
> str(df)
```

```
,data.frame,:   54 obs. of  4 variables:
 $ Year      : int  1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 ...
 $ Population: int  2473000 2539000 2578000 2668000 2771305 2980000 3157000 3310000 3505000 3747000 ...
 $ Attacks   : int  0 1 0 0 1 0 3 1 0 0 ...
 $ Fatalities: int  0 1 0 0 0 0 1 0 0 0 ...
```

```
> rate <- with(df, 1e+06 * Attacks/Population)
> print(xyplot(rate ~ Year, data = df, type = list("g", "p", "smooth"),
     xlab = "Year", ylab = "Rate per Million Pop.", main = "Shark Attacks"))
```



Obrázek 1: Bodový graf dat *Shark Attacks* spolu s odhadem trendu.

Jak již bylo řečeno, místo binomického rozdělení budeme pracovat s Poissonovým rozdělení. Protože jde o podílová data, nesmíme zapomenout na *offset*.

Nejprve začneme s nejjednodušším modelem

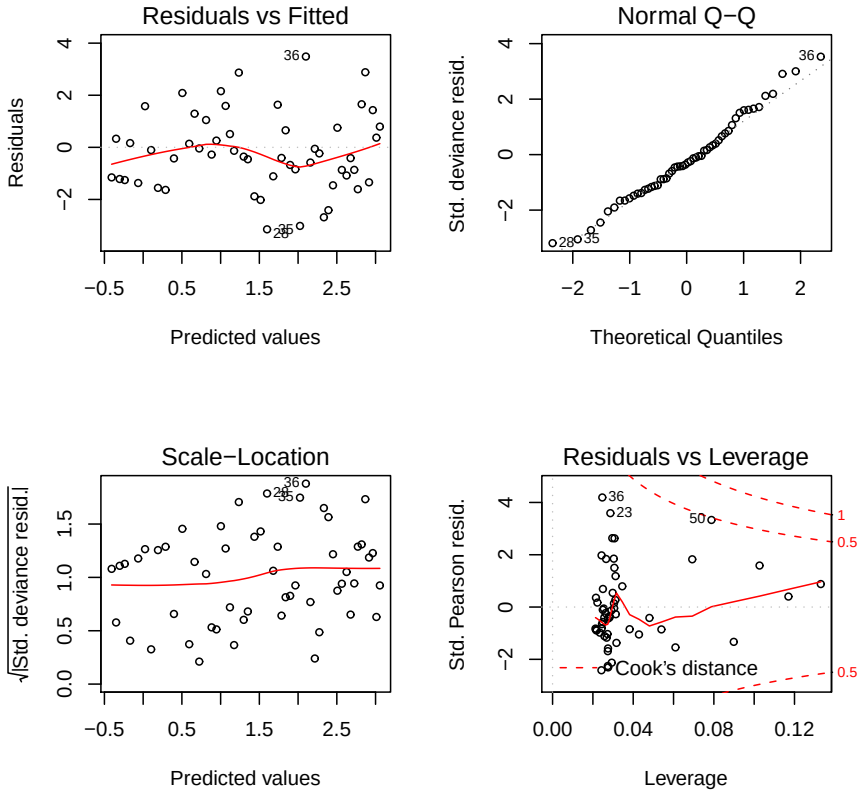
$$\eta_i = \log(\mu_i) = \log(Population) + \beta_0 + \beta_1 x_i \qquad \text{kde} \quad x_i, \quad \text{je proměnná } Year.$$

```
> llfit <- glm(Attacks ~ offset(log(Population)) + Year, data = df, family = poisson)
> summary(llfit)
```

```
Call:
glm(formula = Attacks ~ offset(log(Population)) + Year, family = poisson,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1470  -1.2001  -0.3177   0.7281   3.4856

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -75.792269   8.658221  -8.754  < 2e-16 ***
Year          0.031174   0.004361   7.148  8.8e-13 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 176.93  on 53  degrees of freedom
Residual deviance: 119.11  on 52  degrees of freedom
AIC: 288.76

Number of Fisher Scoring iterations: 5
```

Vidíme, že jednotlivé kovariáty jsou statisticky významné. Nesmíme zapomenout také na anlýzu reziduí.

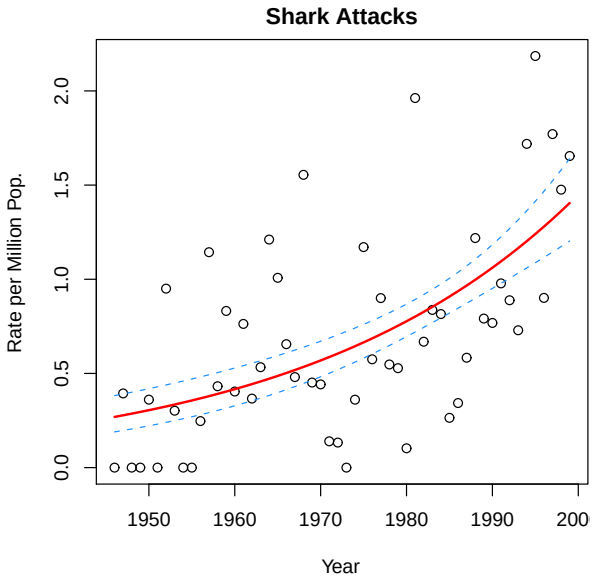
```
> par(mfrow = c(2, 2))
> plot(llfit, cex = 0.75)
```



Obrázek 2: Analýza reziduí pro poissonovskou regresi s lineárních trendem v lineárním prediktoru pro data *Shark Attacks*.

Graficky znázorníme průběh trendu.

```
> Pred <- predict(llfit, type = "response")
> Pred.log <- predict(llfit, type = "link", se = T)
> CI.L.log <- with(df, 1e+06 * exp(Pred.log$fit - 1.96 * Pred.log$se.fit)/Population)
> CI.H.log <- with(df, 1e+06 * exp(Pred.log$fit + 1.96 * Pred.log$se.fit)/Population)
> predRate <- with(df, 1e+06 * Pred/Population)
> par(mfrow = c(1, 1), mar = c(4, 4, 2, 0) + 0.05)
> with(df, plot(Year, rate, xlab = "Year", ylab = "Rate per Million Pop.",
    main = "Shark Attacks"))
> with(df, lines(Year, predRate, lty = 1, col = "red", lwd = 2))
> with(df, lines(Year, CI.L.log, lty = 2, col = "dodgerblue", lwd = 1))
> with(df, lines(Year, CI.H.log, lty = 2, col = "dodgerblue", lwd = 1))
```



Obrázek 3: Poissonovská regrese s lineárních trendem v lineárním prediktoru pro data *Shark Attacks* spolu s odhadem trendu a intervalu spolehlivosti.

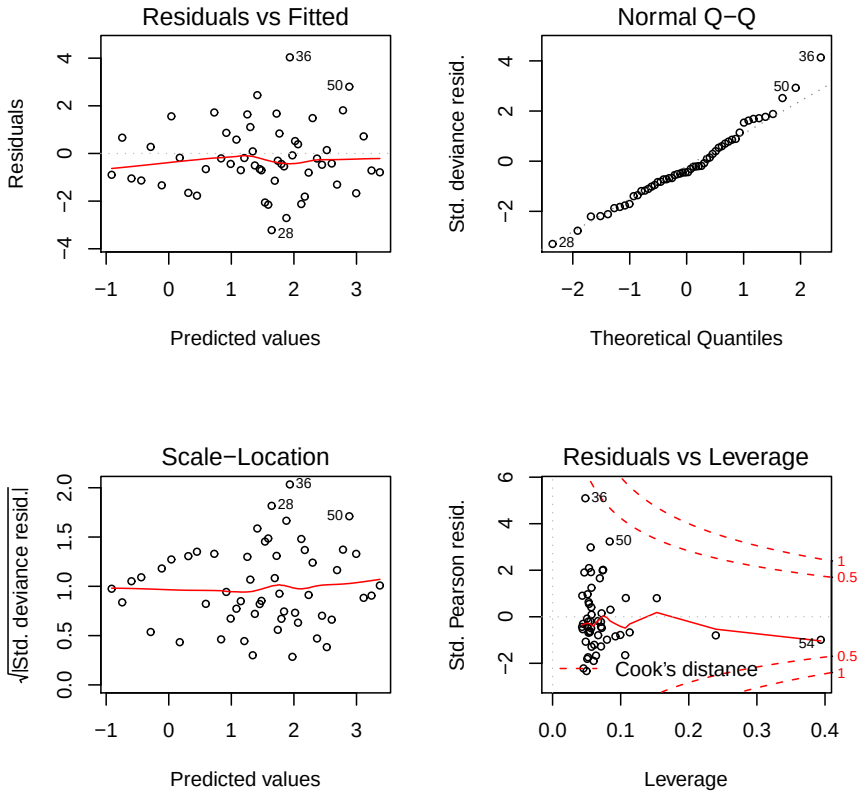
Vzhledem k neparametrickému odhadu trendu v lineárním predikoru nevystačíme s lineárním vztahem, proto vyzkoušíme polynom třetího řádu.

$\eta_i = \log(\mu_i) = \log(Population) + \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3,$ kde $x_i,$ je proměnná *Year*.

```
> llfit3 <- glm(Attacks ~ offset(log(Population)) + Year + I(Year^2) +
     I(Year^3), data = df, family = poisson)
> summary(llfit3)


Call:
glm(formula = Attacks ~ offset(log(Population)) + Year + I(Year^2) +
    I(Year^3), family = poisson, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2177  -1.0113  -0.4306   0.6417   4.0344

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.894e+05  1.775e+05  -2.757  0.00583 **
Year         7.439e+02  2.694e+02   2.762  0.00575 **
I(Year^2)   -3.769e-01  1.362e-01  -2.766  0.00567 **
I(Year^3)    6.365e-05  2.297e-05   2.771  0.00558 **
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 176.93  on 53  degrees of freedom
Residual deviance: 107.18  on 50  degrees of freedom
AIC: 280.83

Number of Fisher Scoring iterations: 5
```

Vidíme, že jednotlivé kovariáty jsou statisticky významné. Nesmíme zapomenout také na anlýzu reziduí.

```
> par(mfrow = c(2, 2))
> plot(llfit3, cex = 0.75)
```



Obrázek 4: Analýza reziduí pro poissonovskou regresi s polynomem 3. řádu v lineárním prediktoru pro data *Shark Attacks*.

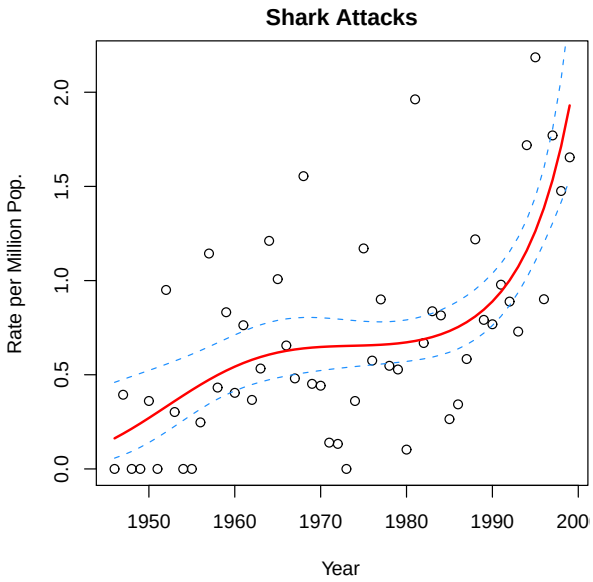
Oba dva modely ihned porovnáme.

```
> anova(llfit3, llfit, test = "Chi")

Analysis of Deviance Table

Model 1: Attacks ~ offset(log(Population)) + Year + I(Year^2) + I(Year^3)
Model 2: Attacks ~ offset(log(Population)) + Year
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        50     107.18
2        52     119.11 -2  -11.932  0.002564 **
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1
```

Vidíme, že došlo k významnému zlepšení modelu.

Graficky znázorníme průběh trendu pro model, který má v lineárním prediktoru polynom 3. řádu.

```
> Pred <- predict(llfit3, type = "response")
> Pred.log <- predict(llfit3, type = "link", se = T)
> CI.L.log <- with(df, 1e+06 * exp(Pred.log$fit - 1.96 * Pred.log$se.fit)/Population)
> CI.H.log <- with(df, 1e+06 * exp(Pred.log$fit + 1.96 * Pred.log$se.fit)/Population)
> predRate <- with(df, 1e+06 * Pred/Population)
> par(mfrow = c(1, 1), mar = c(4, 4, 2, 0) + 0.05)
> with(df, plot(Year, rate, xlab = "Year", ylab = "Rate per Million Pop.",
    main = "Shark Attacks"))
> with(df, lines(Year, predRate, lty = 1, col = "red", lwd = 2))
> with(df, lines(Year, CI.L.log, lty = 2, col = "dodgerblue", lwd = 1))
> with(df, lines(Year, CI.H.log, lty = 2, col = "dodgerblue", lwd = 1))
```



Obrázek 5: Poissonovská regrese s polynomem 3. řádu v lineárním prediktoru pro data *Shark Attacks* spolu s odhadem trendu a intervalu spolehlivosti.

Když se zpětně podíváme na vztah mezi stupněmi volnosti a reziduální deviací u obou modelů

```
> (Scale1 <- llfit$deviance/llfit$df.residual)
```

```
[1] 2.290564
```

```
> (Scale3 <- llfit3$deviance/llfit3$df.residual)
```

```
[1] 2.143544
```

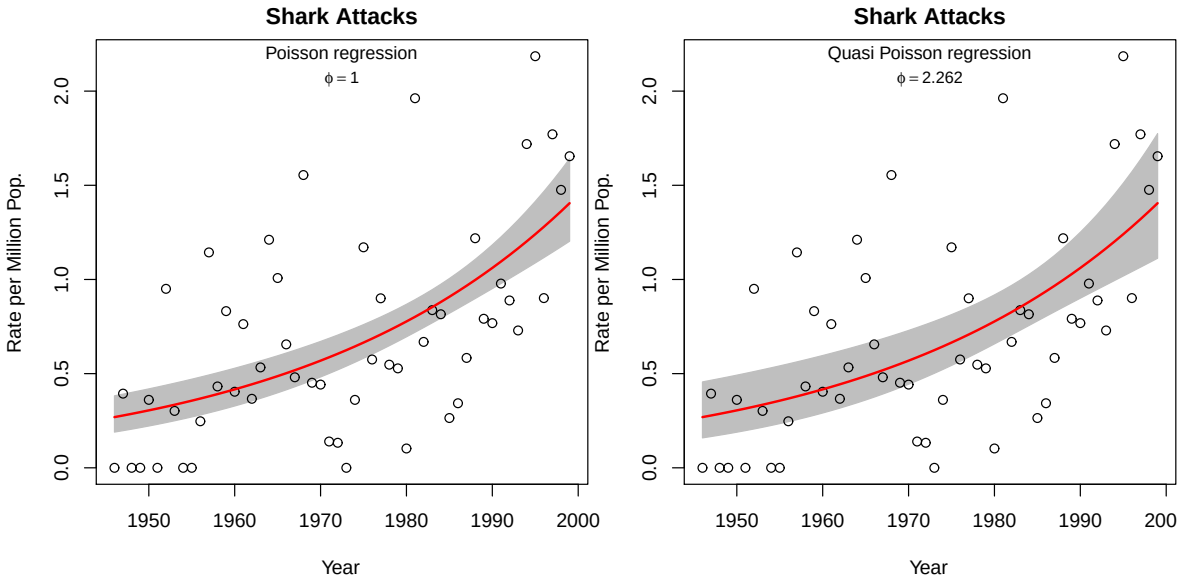
vidíme, že obě hodnoty jsou zhruba dvakrát vyšší. Zvolíme tedy znovu pro oba modelu variantu s volbou

`family=quasipoisson`.

```
> llfitq <- glm(Attacks ~ offset(log(Population)) + Year, data = df, family = quasipoisson)
> summary(llfitq)


Call:
glm(formula = Attacks ~ offset(log(Population)) + Year, family = quasipoisson,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1470  -1.2001  -0.3177   0.7281   3.4856

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -75.792269  13.021371  -5.821 3.69e-07 ***
Year          0.031174   0.006559   4.753 1.62e-05 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

(Dispersion parameter for quasipoisson family taken to be 2.26181)

    Null deviance: 176.93  on 53  degrees of freedom
Residual deviance: 119.11  on 52  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5


> llfit3q <- glm(Attacks ~ offset(log(Population)) + Year + I(Year^2) +
     I(Year^3), data = df, family = quasipoisson)
> summary(llfit3q)


Call:
glm(formula = Attacks ~ offset(log(Population)) + Year + I(Year^2) +
    I(Year^3), family = quasipoisson, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2177  -1.0113  -0.4306   0.6417   4.0344

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.894e+05  2.585e+05  -1.893   0.0641 .
Year         7.439e+02  3.923e+02   1.896   0.0637 .
I(Year^2)   -3.769e-01  1.984e-01  -1.900   0.0633 .
I(Year^3)    6.365e-05  3.345e-05   1.903   0.0628 .
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

(Dispersion parameter for quasipoisson family taken to be 2.121029)

    Null deviance: 176.93  on 53  degrees of freedom
Residual deviance: 107.18  on 50  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

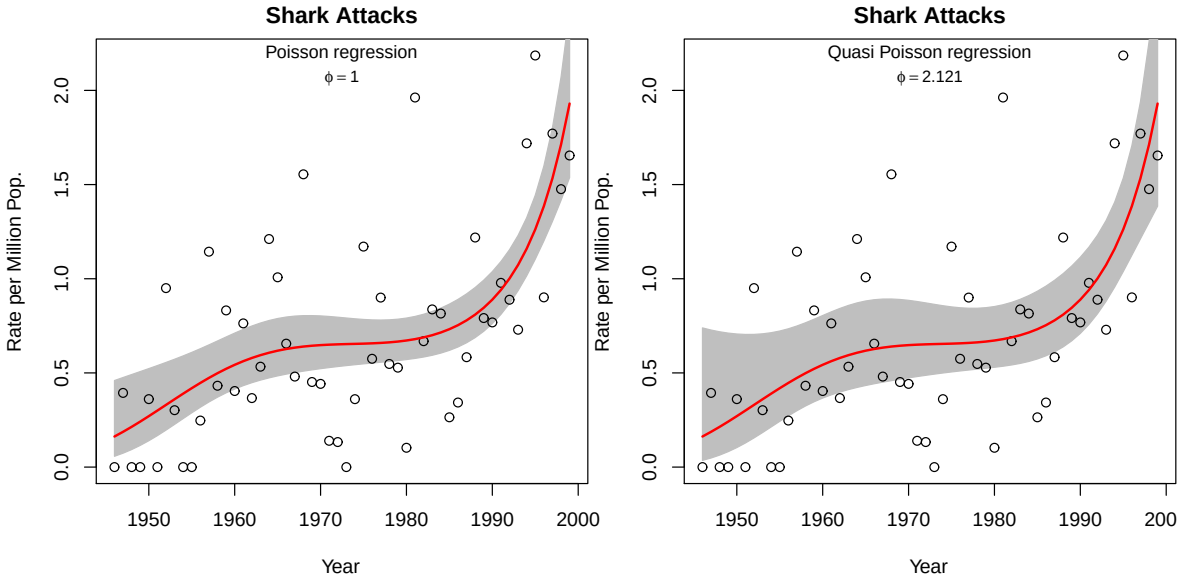
Změnu ve variabilitě odhadů pro oba dva modely graficky porovnáme.

```
> x <- with(df, c(Year, rev(Year)))
> par(mfrow = c(1, 2), mar = c(4, 4, 2, 0) + 0.05)
> Pred <- predict(llfit, type = "response")
> Pred.log <- predict(llfit, type = "link", se = T)
> CI.L.log <- with(df, 1e+06 * exp(Pred.log$fit - 1.96 * Pred.log$se.fit)/Population)
> CI.H.log <- with(df, 1e+06 * exp(Pred.log$fit + 1.96 * Pred.log$se.fit)/Population)
> predRate <- with(df, 1e+06 * Pred/Population)
> y <- c(CI.L.log, rev(CI.H.log))
> with(df, plot(Year, rate, type = "n", xlab = "Year", ylab = "Rate per Million Pop.",
    main = "Shark Attacks"))
> with(df, polygon(x, y, col = "gray75", border = "gray75"))
> with(df, points(Year, rate))
> with(df, lines(Year, predRate, lty = 1, col = "red", lwd = 2))
> mtext("Poisson regression", line = -1, cex = 0.95)
> phi <- round(summary(llfit)$dispersion, 3)
> mtext(text = bquote(phi == .(phi)), side = 3, line = -2, cex = 0.85)
> Pred <- predict(llfitq, type = "response")
> Pred.log <- predict(llfitq, type = "link", se = T)
> CI.L.log <- with(df, 1e+06 * exp(Pred.log$fit - 1.96 * Pred.log$se.fit)/Population)
> CI.H.log <- with(df, 1e+06 * exp(Pred.log$fit + 1.96 * Pred.log$se.fit)/Population)
> predRate <- with(df, 1e+06 * Pred/Population)
> y <- c(CI.L.log, rev(CI.H.log))
> with(df, plot(Year, rate, type = "n", xlab = "Year", ylab = "Rate per Million Pop.",
    main = "Shark Attacks"))
> with(df, polygon(x, y, col = "gray75", border = "gray75"))
> with(df, points(Year, rate))
> with(df, lines(Year, predRate, lty = 1, col = "red", lwd = 2))
> mtext("Quasi Poisson regression", line = -1, cex = 0.95)
> phi <- round(summary(llfitq)$dispersion, 3)
> mtext(text = bquote(phi == .(phi)), side = 3, line = -2, cex = 0.85)
```



Obrázek 6: Poissonovská regrese s lineárních trendem v lineárním prediktoru pro data *Shark Attacks* spolu s odhadem trendu a intervalu spolehlivosti – bez a s řešením problému *over-dispersion*.

```
> x <- with(df, c(Year, rev(Year)))
> par(mfrow = c(1, 2), mar = c(4, 4, 2, 0) + 0.05)
> Pred <- predict(llfit3, type = "response")
> Pred.log <- predict(llfit3, type = "link", se = T)
> CI.L.log <- with(df, 1e+06 * exp(Pred.log$fit - 1.96 * Pred.log$se.fit)/Population)
> CI.H.log <- with(df, 1e+06 * exp(Pred.log$fit + 1.96 * Pred.log$se.fit)/Population)
> predRate <- with(df, 1e+06 * Pred/Population)
> y <- c(CI.L.log, rev(CI.H.log))
> with(df, plot(Year, rate, type = "n", xlab = "Year", ylab = "Rate per Million Pop.",
    main = "Shark Attacks"))
> with(df, polygon(x, y, col = "gray75", border = "gray75"))
> with(df, points(Year, rate))
> with(df, lines(Year, predRate, lty = 1, col = "red", lwd = 2))
> mtext("Poisson regression", line = -1, cex = 0.95)
> phi <- round(summary(llfit3)$dispersion, 3)
> mtext(text = bquote(phi == .(phi)), side = 3, line = -2, cex = 0.85)
> Pred <- predict(llfit3q, type = "response")
> Pred.log <- predict(llfit3q, type = "link", se = T)
> CI.L.log <- with(df, 1e+06 * exp(Pred.log$fit - 1.96 * Pred.log$se.fit)/Population)
> CI.H.log <- with(df, 1e+06 * exp(Pred.log$fit + 1.96 * Pred.log$se.fit)/Population)
> predRate <- with(df, 1e+06 * Pred/Population)
> y <- c(CI.L.log, rev(CI.H.log))
> with(df, plot(Year, rate, type = "n", xlab = "Year", ylab = "Rate per Million Pop.",
    main = "Shark Attacks"))
> with(df, polygon(x, y, col = "gray75", border = "gray75"))
> with(df, points(Year, rate))
> with(df, lines(Year, predRate, lty = 1, col = "red", lwd = 2))
> mtext("Quasi Poisson regression", line = -1, cex = 0.95)
> phi <- round(summary(llfit3q)$dispersion, 3)
> mtext(text = bquote(phi == .(phi)), side = 3, line = -2, cex = 0.85)
```



Obrázek 7: Poissonovská regrese s polynomem 3. řádu v lineárním prediktoru pro data *Shark Attacks* spolu s odhadem trendu a intervalu spolehlivosti – bez a s řešením problému *over-dispersion*.