Štatistické metódy - Projekt

Róbert Kendereš

9. februára 2025

1 Hodnotenia albumov

V prvej časti sme analyzovali dataset obsahujúci hodnotenia 1265 albumov od kritika Anthonyho Fantana (TheNeedleDrop), ktorý obsahoval aj iné informácie o hodnotených albumoch ako

```
album, artist, date, genre, name, score, tags, url.
```

Po prečistení dát sme vybrali len relevantné atribúty -

```
album, genre, score.
```

Najprv sme otestovali normalitu atribútu score (ratings):

```
ks.test(ratings, "pnorm", mean = mean(ratings), sd = sd(ratings))
# p-value < 2.2e-16
```

Kolmogorov-Smirnovov test ukázal, že dáta hodnotení niesú normálne, preto sme pri testovaní hypotéz použili Wilcoxonov test.

1.1 Analýza priemerov

1.1.1 Jeden priemer

Pre jeden priemer sme chceli overiť, či sú v priemere hodnotenia albumov vyššie než 5. Testovali sme hypotézu:

$$H_0 : \mu \leq 5 \quad vs \quad H_1 : \mu > 5$$

Wilcoxonov test

```
wilcox.test(ratings, alternative = "greater", mu = 5)
```

zamietol H_0 ($p < 2.2 \cdot 10^{-16}$), čo znamená, že priemerné hodnotenie albumov je väčšie ako 5.

1.1.2 Dva priemery

Z dát sme vybrali 2 najhodnotenejšie žánre a podľa nich vytvorili dve skupiny:

```
hiphop <- albums[grepl("hip hop", albums$genre),]
pop <- albums[grepl("pop", albums$genre),]
```

V datasete bolo o niečo viac hodnotení pre hip-hop, a tak sme očakávali že by Anthony mal preferenciu pre tento žáner a hodnotil ho lepšie.
Zložili sme hypotézu:

$$H_0 : \mu_{hiphop} \leq \mu_{pop} \quad vs \quad H_1 : \mu_{hiphop} > \mu_{pop}$$

Wilcoxonov test nepreukázal štatisticky významný rozdiel ($p = 0.9998$), čo znamená, že nemôžeme tvrdiť, že hip-hop má vyššie priemerné skóre ako pop.

2 Platy futbalistov

Pri korelačnej analýze sme využili dataset pozostávajúci z 32 riadkov s atribútmi

```
Team, Wins, Avg Age, Active, Salary(total in $M).
```

Z nich nás zaujímali len

- **Wins** - počet výhier tímu za sezónu
- **Avg Age** - priemerný vek hráčov v danom tíme
- **Salary** - celkový plat futbalistov vo vybraných tímoch

2.1 Korelačná analýza

Testovali sme 2 páry hypotéz:

2.1.1 Jeden korelačný koeficient

Obojstranný test o korelácii medzi počtom výhier $Wins$ a priemerným vekom hráčov Age:

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$$

Použitím

```
cor.test(Wins, Age, alternative = "two.sided")
```

sme dostali p-value 0.05346, čiže nemáme dostatočné informácie na zamietnutie nulovej hypotézy.
Ajkeď $r \doteq 0.35$, nieje štatisticky významná na hladine významnosti 5%.
Vypočítali sme Fisherovu Z premennú a pomocou nej zostrojili 95%-ný interval spoľahlivosti pre ρ:

$$Z = \tanh^{-1}(r) = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

$$L = Z - \frac{q_{0.975}}{\sqrt{n-3}}, \quad U = Z + \frac{q_{0.975}}{\sqrt{n-3}}$$

$$(-0.0047,\ 0.7232)$$

Interval hovorí, že skutočná hodnota ρ sa s 95% pravdepodobnosťou nachádza medzi -0.0047 a 0.7232.

2.1.2 Rovnosť dvoch korelačných koeficientov

Ako ďalší sme urobili obojstranný test o rovnosti korelácií medzi počtom výhier a priemerným vekom hráčov pre tímy s platom väčším a menším než je priemerný plat:

$$H_0 : \rho_1 = \rho_2 \quad vs \quad H_1 : \rho_1 \neq \rho_2$$

kde ρ_1 je skutočný korelačný koeficient pre skupinu s nadpriemerným platom a ρ_2 je korelačný koeficient skupiny s platom menším ako priemer.
Najskôr sme rozdelili dáta o veku a výhrach do dvoch skupín na základe priemerného platu (μ - priemer stĺpca Salary):

- Nadpriemerný plat (Salary $\geq \mu$):

```
aboveAge<-Age[Salary>=mu]
aboveWins <- Wins[Salary>=mu]
```

- Podpriemerný plat (Salary $< \mu$):

```
belowAge <- Age[Salary < mu]
belowWins <- Wins[Salary < mu]
```

Pre každú skupinu sme vypočítali korelačné koeficienty a aplikovali Fisher transformáciu:

$$r_1 = \mathrm{cor}(\mathrm{Age}_{\mathrm{above}}, \mathrm{Wins}_{\mathrm{above}}), \quad r_2 = \mathrm{cor}(\mathrm{Age}_{\mathrm{below}}, \mathrm{Wins}_{\mathrm{below}})$$
$$Z_1 = \tanh^{-1}(r_1), \quad Z_2 = \tanh^{-1}(r_2)$$

A získali testovaciu štatistiku

$$Z_{12} = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

kde n_1 a n_2 sú veľkosti daných skupín.
Výsledná p-value testu bola:

$$p = 2 \cdot P(Z > |Z_{12}|) = 0.2593,$$

čo hovorí, že nulovú hypotézu nezamietame, a teda že korelácie medzi vekom hráčov a počtom výhier sa medzi danými skupinami signifikantne nelíšia.

3 Tvrdosť vody

V poslednej časti sme využili dataset s informáciami o obsahu vápnika vo vode a ročnej úmrtnosti v rôznych častiach Anglicka a Walesu. V dátach bolo 61 záznamov (pre 61 rôznych miest) s atribútmi

```
Mortality, Calcium, Region,
```

kde

- **Mortality** - ročná úmrtnosť na 100 000 obyvateľov
- **Calcium** - koncentrácia vápnika v pitnej vode v ppm(parts per million)
- **Region** - (North/South) región kde sa nachádza mesto danej vzorky

Cieľom analýzy bolo zistiť, či existuje štatisticky významný vzťah medzi tvrdosťou vody a úmrtnosťou a ako dobre vieme úmrtnosť predpovedať podľa tvrdosti vody.

3.1 Lineárna regresia

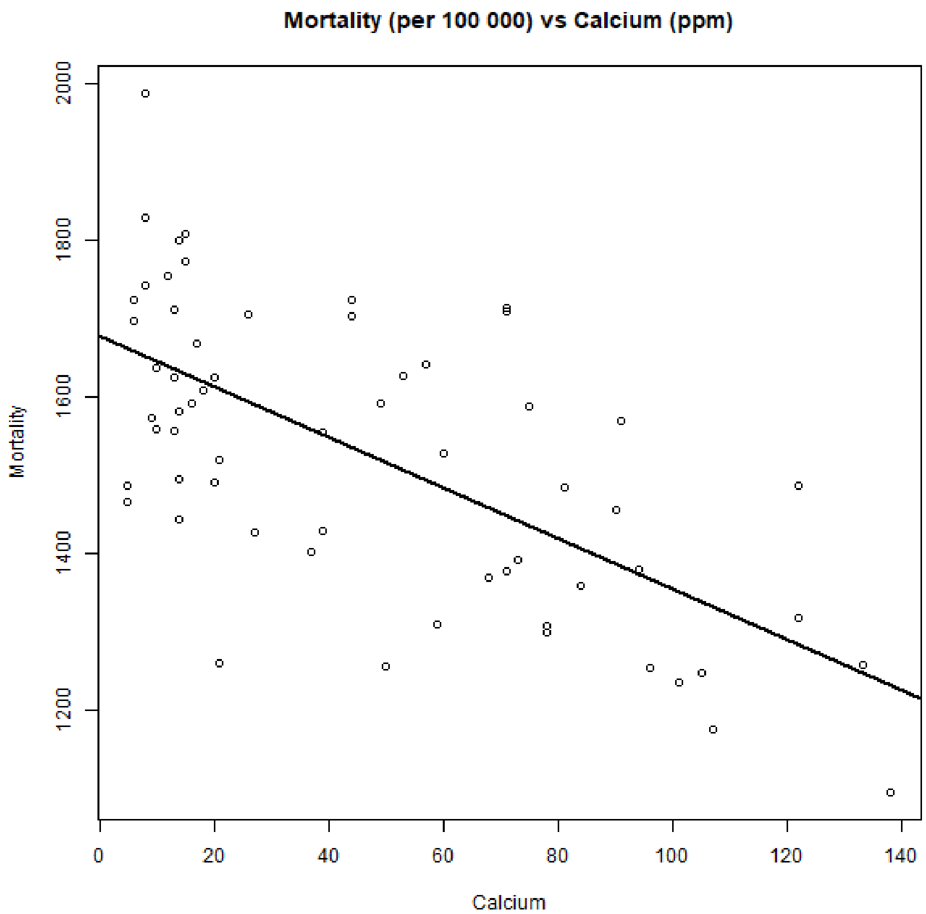
Na začiatok sme vytvorili jednoduchý lineárny regresný model, kde závislou premennou je úmrtnosť (*Mortality*) a nezávislou premennou je tvrdosť vody (*Calcium*):

$$\mathrm{Mortality} = \beta_0 + \beta_1 \cdot \mathrm{Calcium} + \varepsilon.$$

Odhadli sme **parametre modelu**:

$$\mathrm{Mortality} = 1676.36 - 3.23 \cdot \mathrm{Calcium}.$$

- Intercept $\hat{\beta}_0 = 1676.36$ predstavuje očakávanú úmrtnosť pri nulovej tvrdosti vody.
- Koeficient $\hat{\beta}_1 = -3.23$ - každé zvýšenie obsahu vápnika o 1 ppm korešponduje poklesu úmrtnosti o 3.23.



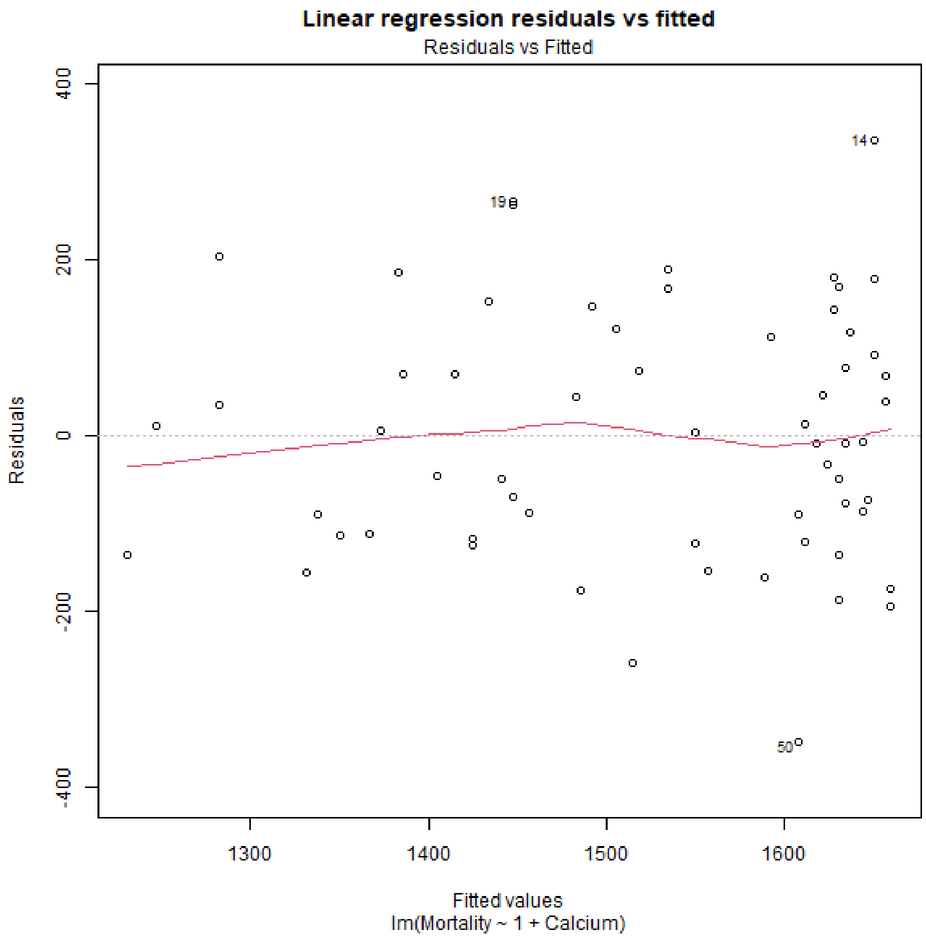
Obr. 1: Regresná priamka pre lineárny model tvrdosti vody a úmrtnosti

Vykonali sme aj **testy normality** pre závislé a nezávislé premenné

```
ks.test(Mortality, "pnorm", mean = mean(Mortality), sd = sd(Mortality))
#p-value = 0.8968
ks.test(Calcium, "pnorm", mean = mean(Calcium), sd = sd(Calcium))
#p-value = 0.01786,
```

kde sa ukázalo, že Calcium nieje normálne rozdelené.
Podmienka regresie je ale normalita reziduí, ktorú sme Shapiro-Wilk testom potvrdili s p-value = 0.6798.

```
shapiro.test(MODEL$residuals)     #p-value = 0.6798
```



Obr. 2: Graf rezíduí pre lineárny model, približne rovnomerné rozdelenie po vodorovnej osi.

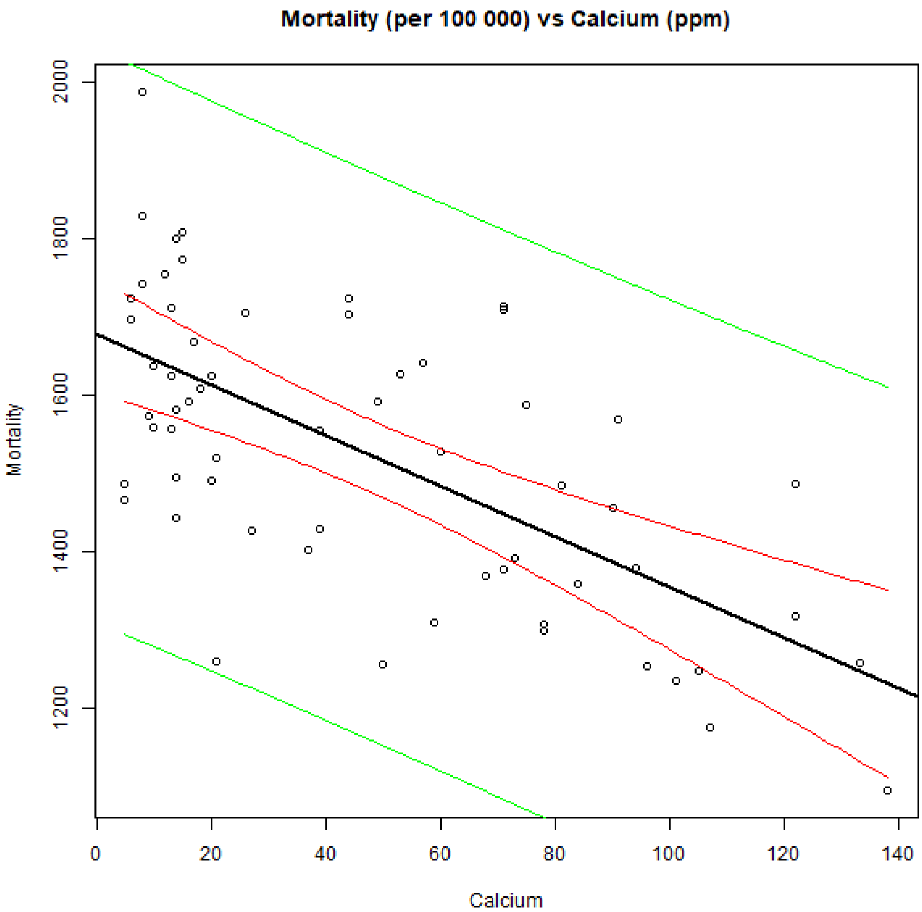
Štatistická významnosť modelu:

- F-test: $p = 1.033 \cdot 10^{-8} < 0.05$ indikuje, že model má signifikantný predikčný výkon.
- koeficient determinácie $R^2 = 0.4288$ ale ukázal, že tvrdosť vody vysvetľuje iba približne 42.9% variability úmrtnosti čo znamená že veľa dát nevie popísať veľmi dôveryhodne.

3.1.1 Pás spoľahlivosti a predikčný pás

Na obrázku 3 sú zakreslené pás spoľahlivosti(červený) a predikčný pás(zelený) pre regresnú priamku:

- **Čierna** : Odhadovanú regresná priamka.
- **Červené čiary:** Ohraničujú oblasť 95%-ného intervalu spoľahlivosti pre očakávanú hodnotu úmrtnosti $\mathbb{E}$(Mortality), teda vyjadruje rozsah hodnôt, v ktorom s 95%-nou pravdepodobnosťou leží skutočná očakávaná hodnota Mortality pre nejakú hodnotu Calcium.
- **Širší pás ohraničený zelenými čiarami:** Reprezentuje 95%-ný predikčný interval - ukazuje rozsah hodnôt, v ktorých sa očakáva, že sa bude nachádzať nová, individuálna pozorovaná hodnota Mortality pri určitej hodnote Calcium.



Obr. 3: Lineárna regresia s pásom spoľahlivosti a predikčným pásom

Hranice pre dané pásy sme vypočítali Scheffeho metódou:

$$IS : \mathbf{a}^T\hat{\boldsymbol{\beta}} \pm \sqrt{2 \cdot F_{0.95,2,n-2}} \cdot s \cdot \sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}},$$

$$PI : \mathbf{a}^T\hat{\boldsymbol{\beta}} \pm \sqrt{2 \cdot F_{0.95,2,n-2}} \cdot s \cdot \sqrt{1 + \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}},$$

kde

- $\mathbf{a}$ je vektor tvaru $(1, x)^T$,
- $\hat{\boldsymbol{\beta}}$ je vektor odhadnutých parametrov,
- s je reziduálna smer. odchýlka, $s = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-k}}$, kde k je počet parametrov modelu (2),
- $F_{0.95,2,n-2}$ je kvantil Fisherovho rozdelenia s $\alpha = 5\%$ a 2, n-2 stupňami voľnosti,
- $\mathbf{X}$ je matica nezávislých premenných

3.1.2 IS a PI pre kontrast

Podobne sme vypočítali IS a PI pre kontrast:

$$IS : \mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}},$$

$$PI : \mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}},$$

kde $t_{\alpha/2,n-2}$ je kritická hodnota t=rozdelenia s n-2 stupňami voľnosti na $\alpha = 5\%$ a $\mathbf{a}$ je kontrastný vektor. Pre $\mathbf{a} = (1, -1)^T$ nám vyšli hodnoty:

- IS: (1620.196, 1738.967)
- PI: (1387.284, 1971.879)
- $\mathbf{a}^T\hat{\boldsymbol{\beta}} = 1679.582$

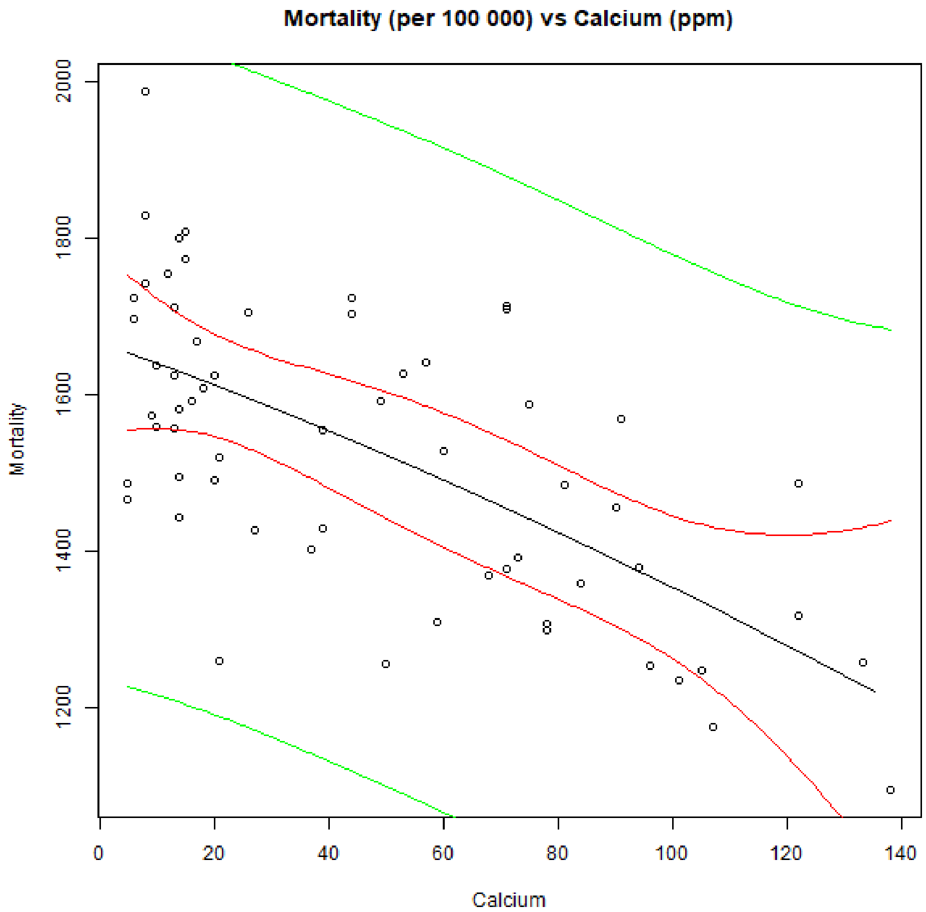
Je hneď vidieť, podobne ako na obrázku, že predikčný interval je omnoho širší ako interval spoľahlivosti. Je to tak kvôli tomu, že IS vyjadruje rozmedzie pre skutočnú **strednú hodnotu** Mortality, nie špecifickú hodnotu pre danú hodnotu Calcium.

3.2 Polynomiálna regresia

Vykonali sme aj polynomiálnu regresiu (kvadratickú) na dátach, aby sme overili, či neopisuje dáta lepšie.

$$\text{Mortality} = \beta_0 + \beta_1 \cdot \text{Calcium} + \beta_2 \cdot \text{Calcium}^2 + \varepsilon.$$

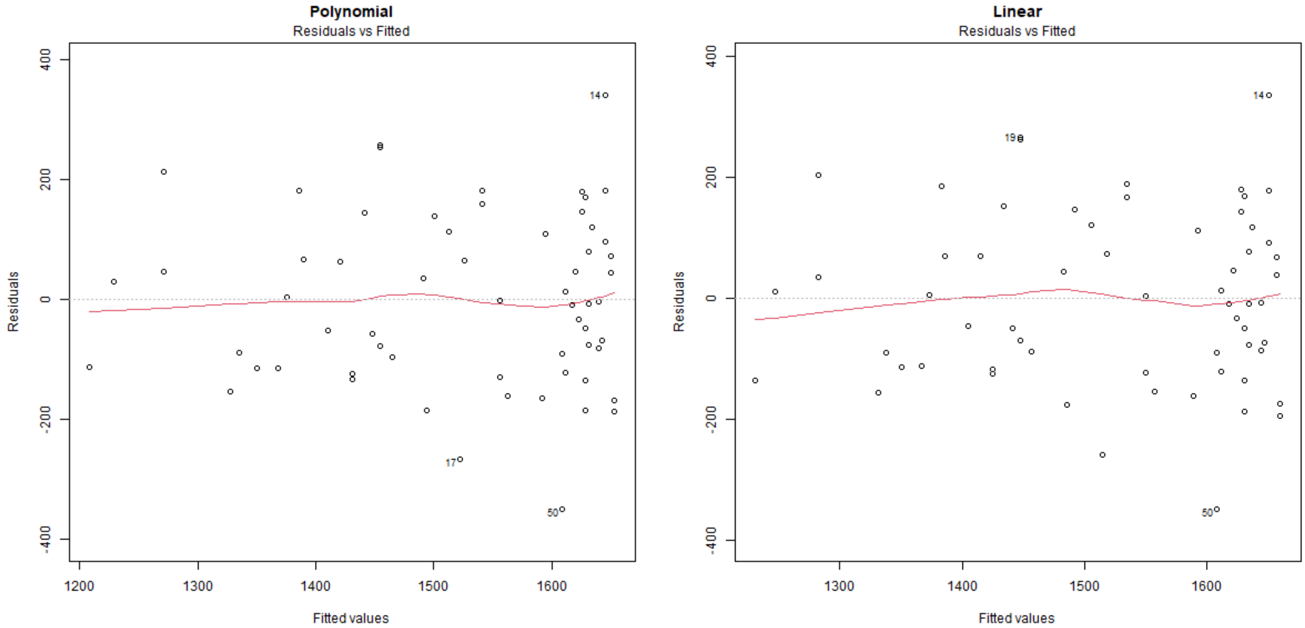
Výsledný model dosiahol mierne vyššie $R^2 = 0.4301$, čo je ale zanedbateľné zlepšenie v porovnaní s lineárnym modelom.



Obr. 4: Polynomiálna regresia s pásom spoľahlivosti a predikčným pásom

3.3 Porovnanie lineárneho a kvadratického modelu

Oba modely poskytujú podobné výsledky a vzťah medzi tvrdosťou vody a úmrtnosťou je približne lineárny, a daný lineárny model by mohol byť do istej miery vhodný na predpovedanie hodnôt úmrtnosti.
Porovnanie reziduí pre oba modely je vykreslené na obrázku 5.



Obr. 5: Porovnanie lineárnej a polynomiálnej regresie ukazuje malý rozdiel v shopnostiach modelov