Kovalčíková Petra
371852

# Regresní modelování – projekt

## I. POPIS DATOVÉHO SOUBORU

Datový soubor je možné nalézt na: [http://www.statsci.org/data/general/fev.txt](http://www.statsci.org/data/general/fev.txt), zdroj dat se odkazuje na publikaci: Tager, I. B., Weiss, S. T., Rosner, B., and Speizer, F. E. (1979). Effect of parental cigarette smoking on pulmonary function in children. *American Journal of Epidemiology*, **110**, 15-26. Jedná se o děti a mládež ve věku 3-19 let (n=654), u kterých byly zjišťovány dýchací funkce prostřednictvím FEV – jednovteřinové vitální kapacity plic (maximální množství vydechnuté za jednu vteřinu, v litrech). O subjektech pak byly zaznamenávány další charakteristiky: pohlaví, věk, výška a kouření (nekuřák/současný kuřák). Díky charakteristikám, které se v datech vyskytují, byla vybrána následující hypotéza: *„Ovlivňuje kouření u dětí a mládeže dýchací funkce měřené pomocí FEV?“* Cílem je vytvořit lineární model pro FEV s využitím proměnných dostupných v souboru dat. Protože se v datovém souboru vyskytují kuřáci pouze ve věku 9 a více let, byl datový soubor omezen pouze na děti a mládež ve věku 9-19 let (n=439) a veškerá analýza dále popsaná je provedena na takto věkově omezeném souboru.

Proměnné datového souboru z odkazu výše (některé proměnné pak byly následně převedeny na jiné jednotky či byly transformovány):

| | | |
|---|---|---|
| *FEV* | jednovteřinová vitální kapacita plic | [l] |
| *Age* | věk | [roky] |
| *Height* | výška | [palce] |
| *Sex (Female/ Male)* | pohlaví (ženy/ muži) | |
| *Smoker (Current/Non)* | kouření (současný kuřák/ nekuřák) | |

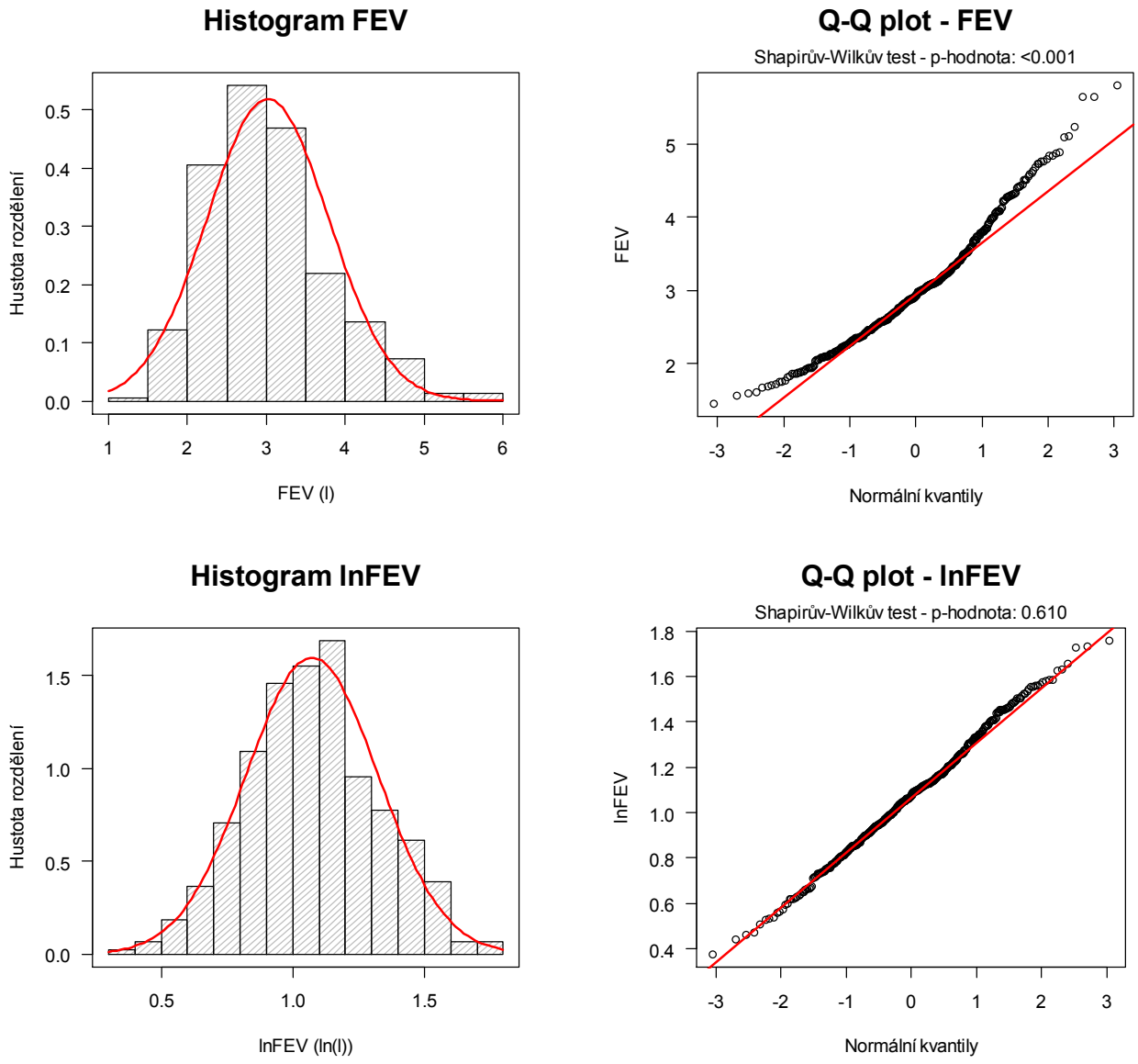
## II. ZÁKLADNÍ POPISNÁ STATISTIKA SOUBORU A TRANSFORMACE DAT

Tabulka 1 ukazuje základní popisnou statistiku souboru. Výška dětí byla z palců převedena na centimetry (*Heightcm*), v tabulce jsou pro ukázku sumarizovány obě varianty. Dále byla provedena transformace FEV přirozeným logaritmem (*lnFEV*), protože původní proměnná se neřídí normálním rozdělením (histogram, normální Q-Q plot a p-hodnota Shapirova-Wilkova testu pro FEV i lnFEV jsou zobrazeny na obrázku 1). Po logaritmizaci FEV již nemůžeme zamítnout nulovou hypotézu o normalitě dat pomocí Shapirova-Wilkova testu (p-hodnota 0.610).

**Tabulka 1.** Základní charakteristika souboru.

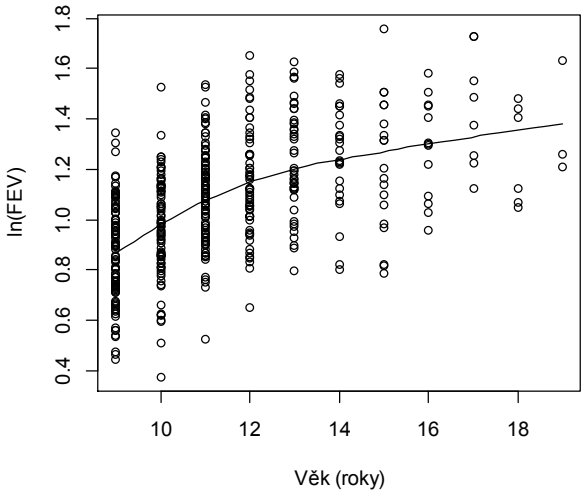
| Charakteristika | Průměr (min-max) nebo n (%) |
|---|---|
| FEV (l) | 3.0 (1.5-5.8) |
| lnFEV (ln(l)) | 1.1 (0.4-1.8) |
| Věk (roky) | 11 (9-19) |
| Výška (palce) | 64 (53-74) |
| Výška (cm) | 163 (133-188) |
| Pohlaví - muži | 232 (52.9) |
| Pohlaví - ženy | 207 (47.2) |
| Kouření - nekuřáci | 374 (85.2) |
| Kouření - současní kuřáci | 65 (14.8) |

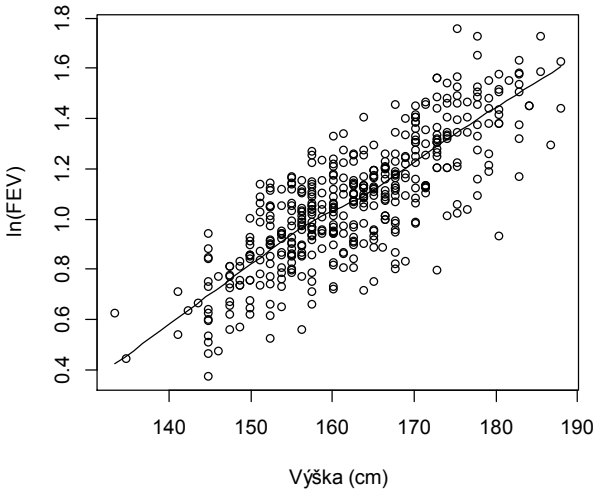
**Obrázek 1.** Hodnocení normality FEV a lnFEV pomocí histogramu a normálního Q-Q plotu.

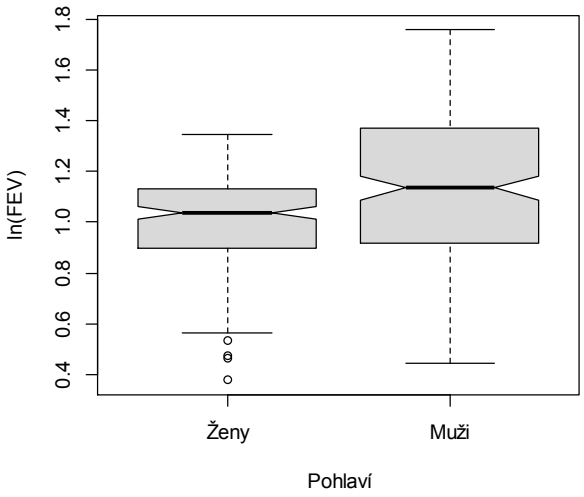


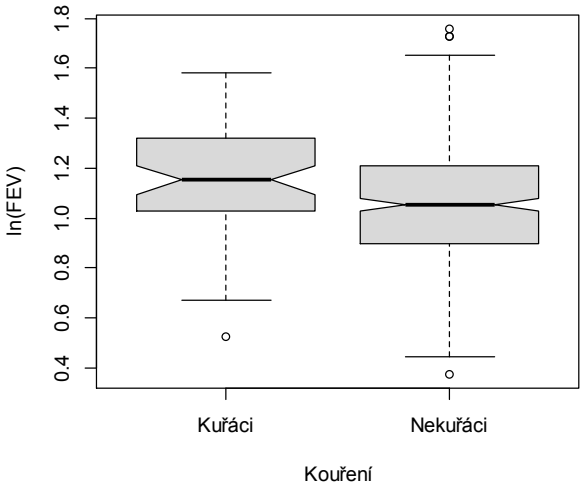

Obrázek 2 ukazuje marginální závislosti jednotlivých proměnných na lnFEV. Z jednorozměrného pohledu se zdá, že výška dětí a mládeže má s lnFEV lineární vztah, věk má také částečně lineární efekt, přičemž je ale možné, že věk a výška dětí a mládeže bude silně korelovat. Dále to vypadá, že chlapci mají lepší dýchací funkce charakterizované FEV než dívky, zatímco nekuřáci horší než kuřáci. Efekty jsou popsány jednorozměrnými lineárními modely v následující kapitole.

**Obrázek 2.** Závislost charakteristik dětí a mládeže na lnFEV.



## III. JEDNOROZMĚRNÁ LINEÁRNÍ REGRESE

V jednorozměrných modelech vyšly všechny proměnné statisticky významně. V tabulce 2 jsou uvedeny výsledky modelů – odhady koeficientů, celková významnost modelu (v tomto případě významnost regresního koeficientu pro danou proměnnou) a index determinace ($R^2$). U pohlaví a kouření je v závorce uvedena referenční skupina (ženy a nekuřáci). Všechny modely vyšly významně, nejvíce variability lnFEV však vysvětluje model s výškou dětí a mládeže. Z jednorozměrného pohledu mají lepší dýchací funkce (lnFEV) starší a vyšší děti, chlapci a kuřáci.

**Tabulka 2.** Výsledky jednorozměrných lineárních modelů závislosti lnFEV na charakteristikách dětí a mládeže.

| **Charakteristika** | **$B_0$** (intercept) | **$B_1$** (sklon) | **p-hodnota** | **$R^2$** |
|---|---|---|---|---|
| Věk (roky) | 0.380 | 0.601 | <0.001 | 0.295 |
| Výška (cm) | -2.145 | 0.020 | <0.001 | 0.646 |
| Pohlaví (ref. ženy) | 1.006 | 0.128 | <0.001 | 0.066 |
| Kouření (ref. nekuřáci) | 1.058 | 0.102 | 0.002 | 0.021 |

Před tvořením modelů z více proměnných by bylo vhodné se podívat na vztahy mezi jednotlivými proměnnými – závislost mezi věkem a výškou byla hodnocena pomocí Pearsonova korelačního koeficientu ($r^2$=0.531, p-hodnota: <0.001) a vykazuje kladnou lineární závislost. Nezávislost mezi pohlavím a kouřením byla testována pomocí chí-kvadrát testu (p-hodnota: 0.036) a výsledky ukazují na závislost i mezi těmito proměnnými. Pomocí Mannova-Whitneyho testu byl hodnocen vztah mezi věkem a pohlavím (p-hodnota: 0.792), věkem a kouřením (p-hodnota: <0.001), výškou a pohlavím (p-hodnota: <0.001) a výškou a kouřením (p-hodnota: <0.001). Zdá se, že většina proměnných spolu souvisí a jsou navzájem závislé. U vícerozměrných modelů bude ještě multikolinearita zjišťována pomocí VIF (Variance Inflation Factors).

## IV. VÍCEROZMĚRNÁ LINEÁRNÍ REGRESE

První model byl uvažován bez interakcí, se všemi proměnnými. Výsledky tohoto modelu jsou následující:

**Model 1:**

```
Call:
lm(formula = lnFEV ~ Age + Heightcm + Sex + Smoker, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50839 -0.09001  0.01665  0.09274  0.37255

Coefficients:
                Estimate  Std. Error t value Pr(>|t|)
(Intercept)    -2.005371   0.120871 -16.591  < 2e-16 ***
Age             0.021817   0.003880   5.624 3.35e-08 ***
Heightcm        0.017410   0.000867  20.079  < 2e-16 ***
SexMale         0.010960   0.015018   0.730   0.4659
SmokerCurrent  -0.050125   0.021153  -2.370   0.0182 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1443 on 434 degrees of freedom
Multiple R-squared:  0.6699,      Adjusted R-squared:  0.6668
F-statistic: 220.2 on 4 and 434 DF,  p-value: < 2.2e-16
```

Tento maximální model (s ohledem na dostupná data v souboru) vysvětluje 67 % variability lnFEV a kromě pohlaví jsou všechny proměnné významné, včetně kouření. Další model bude tedy zjednodušen vynecháním proměnné pohlaví.

**Model 2:**

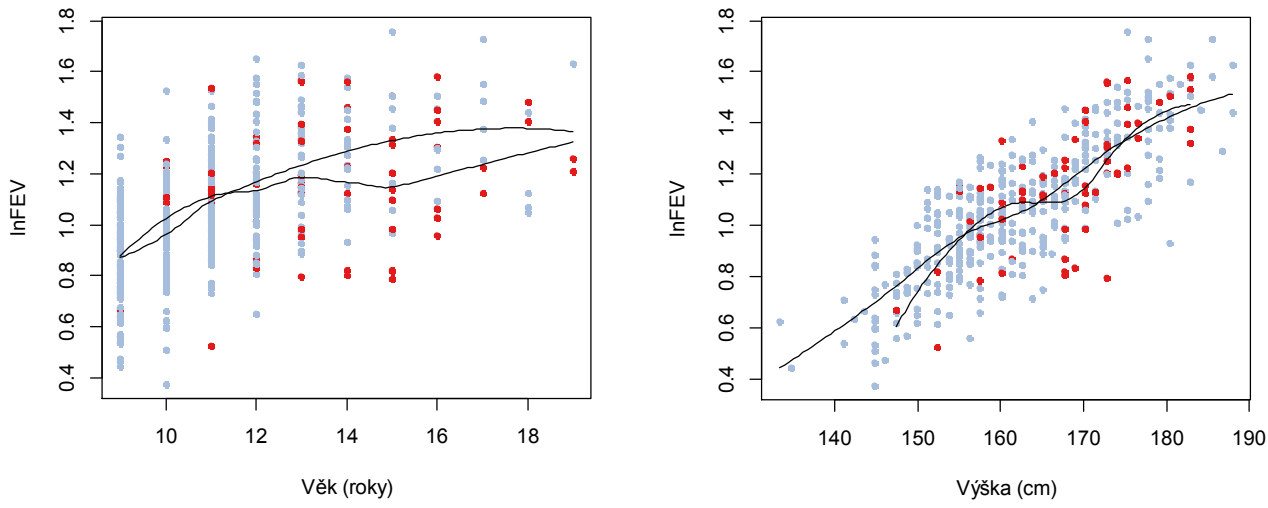
```
Call:
lm(formula = lnFEV ~ Age + Heightcm + Smoker, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51760 -0.08924  0.01325  0.09545  0.37597

Coefficients:
                    Estimate   Std. Error t value Pr(>|t|)
(Intercept)        -2.0331638  0.1146526 -17.733  < 2e-16 ***
Age                 0.0213662  0.0038279   5.582  4.2e-08 ***
Heightcm            0.0176501  0.0008017  22.015  < 2e-16 ***
SmokerCurrent      -0.0520865  0.0209706  -2.484   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1442 on 435 degrees of freedom
Multiple R-squared:  0.6695,      Adjusted R-squared:  0.6672
F-statistic: 293.7 on 3 and 435 DF,  p-value: < 2.2e-16
```

Tento model má adjustované $R^2$ téměř stejné jako v předchozím maximálním modelu (0.667), regresní koeficienty zůstaly téměř nezměněny, všechny proměnné jsou v modelu statisticky významné, stejně tak celý model (F-test). Pomocí ANOVA byla testována hypotéza o vhodnosti jednoduššího modelu, která nebyla zamítnuta (p-hodnota: 0.466) a dále tedy bude pracováno s tímto modelem. Multikolinearita v tomto modelu byla zjišťována pomocí VIF, pro všechny proměnné byly hodnoty VIF menší než 2 (věk: 1.56, výška: 1.39, kouření: 1.17). Nezdá se tedy, že by v tomto modelu byl problém s multikolinearitou.

Tento model ukazuje odlišné výsledky oproti jednorozměrným modelům, protože kouření zde již má negativní efekt na dýchací funkce po adjustaci vlivu věku a výšky dětí a mládeže.

Dále bylo zjišťováno, zda se mezi proměnnými v modelu 2 (tedy všemi s vyloučením pohlaví) vyskytují nějaké významné interakce. Na obrázku 3 jsou pomocí marginálních závislostí vykresleny grafy závislosti lnFEV na věku a výšce s rozlišením kuřáků (červené tečky) a nekuřáků (modré tečky). Grafy ukazují, že by případně mohla být interakce spíše mezi věkem a kouřením než mezi výškou a kouřením, kde to na interakce mezi těmito prediktory nevypadá.

**Obrázek 3.** Marginální závislosti lnFEV na věku a výšce pro kuřáky a nekuřáky.



V dalším modelu tedy byla zařazena i interakce mezi věkem a kouřením.

**Model 3:**

```
Call:
lm(formula = lnFEV ~ Heightcm + Age * Smoker, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52058 -0.08855  0.01208  0.09732  0.36989

Coefficients:
                     Estimate   Std. Error t value Pr(>|t|)
(Intercept)         -2.0283351  0.1147121 -17.682  < 2e-16 ***
Heightcm             0.0174698  0.0008183  21.348  < 2e-16 ***
Age                  0.0235563  0.0043201   5.453 8.34e-08 ***
SmokerCurrent        0.0724252  0.1158601   0.625    0.532
Age:SmokerCurrent   -0.0095210  0.0087132  -1.093    0.275
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1442 on 434 degrees of freedom
Multiple R-squared:  0.6704,      Adjusted R-squared:  0.6673
F-statistic: 220.7 on 4 and 434 DF,  p-value: < 2.2e-16
```

Interakce mezi věkem a kouřením vyšla nevýznamná, zároveň v tomto modelu obsahující interakce i vliv samotného kouření vyšel statisticky nevýznamně. Byly vyzkoušeny i další možné interakce s podobným výsledkem jako v případě věku a kouření.

Jako finální model byl tedy vybrán **model 2** – závislost lnFEV na kouření, věku a výšce dětí a mladistvých. Pro tento model bude v další kapitole provedena analýza reziduí.

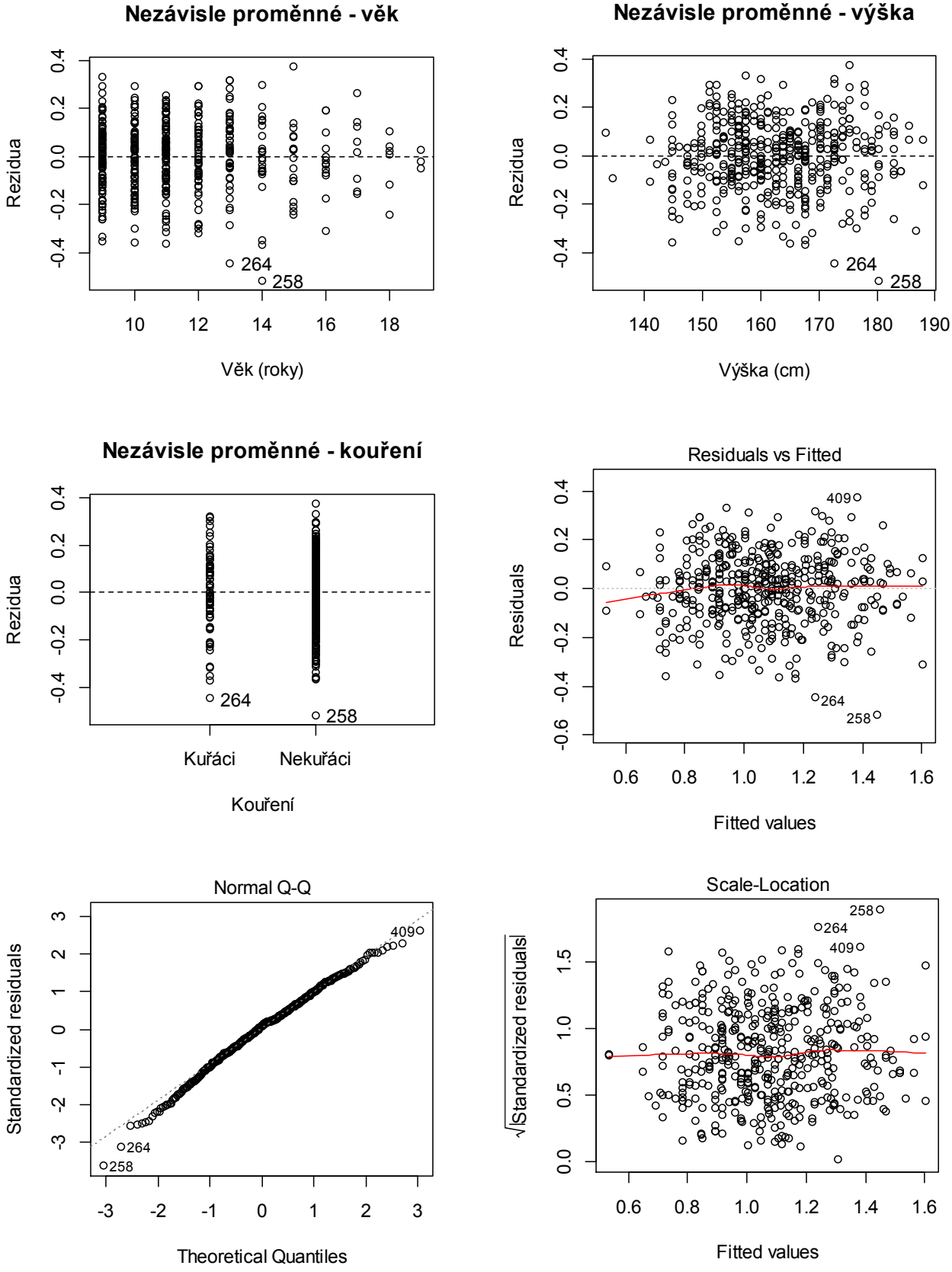
## V. ANALÝZA REZIDUÍ

Obrázek 4 ukazuje hodnocení předpokladů lineárního modelu - grafy reziduí proti nezávislým proměnným, která by měla být symetricky rozptýlena okolo 0, dále pak graf reziduí vůči predikovaným hodnotám lnFEV, normálně-pravděpodobnostní graf reziduí a tzv. Scale-Location plot poskytovaný jako součást analýzy reziduí funkce pro lineární regresní modely v R (*lm*). V grafech reziduí vůči nezávislým proměnným je vypsáno pořadí (ID) dětí, pro nejvyšší hodnoty reziduí v absolutní hodnotě (>0.4) – odlehlé body. Ze všech grafů na obrázku 4 se zdá, že předpoklady modelu (linearita a homogenita rozptylu reziduí, normální rozdělení chyb) nejsou závažně porušeny. Jako případná odlehlá pozorování se zdají být subjekty s ID 258 a 264 dle pořadí v datech.

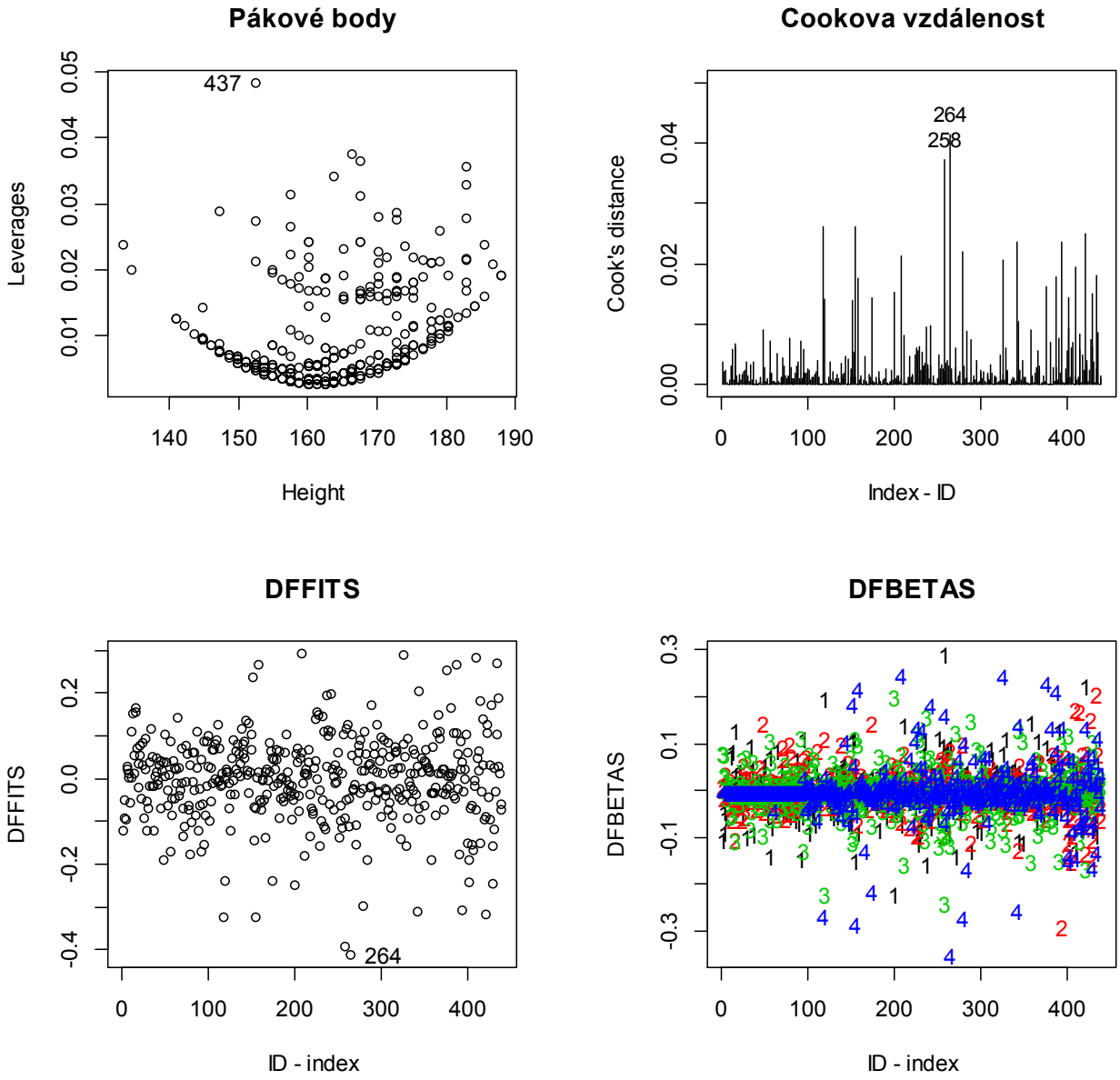
Dále byla hledána vlivná pozorování pomocí Cookovy vzdálenosti, pákových bodů, DFFITS a DFBETAS (pákové body seřazeny dle výšky subjektů, ostatní dle ID podle pořadí). Grafy jsou zobrazeny na obrázku 5. Jako vlivné pozorování byly označeny subjekty s ID 437, 258 a 264.

Nalezená odlehlá a vlivná pozorování by bylo vhodné podrobněji prozkoumat a následně určit, zda se nemůže jednat o chybné hodnoty, v takovém případě by bylo možné je z modelu odstranit a pokračovat v analýze bez nich. Byl proveden stejný model jako finální model (model 2), pouze bez těchto 3 pozorování, ale výsledky modelu (odhady koeficientů, významnost koeficientů a modelu, index determinace) se prakticky nezměnily. Protože se u těchto subjektů nezdá, že by se jednalo o fyziologicky nemožné hodnoty a tedy chybná data, jako finální model byl ponechán model 2 se všemi pozorováními.

**Obrázek 4.** Grafy reziduí proti nezávislým proměnným, reziduí proti predikovaným hodnotám, normální Q-Q plot graf pro standardizovaná rezidua a Scale-Location plot.



**Obrázek 5.** Hledání vlivných pozorování – pákové body, Cookova vzdálenost, DFFITS, DFBETAS.




## VI. SHRNUTÍ

Byla provedena analýza závislosti dýchacích funkcí (charakterizované pomocí FEV – jednovteřinové vitální kapacity plic) u dětí a mládeže na jejich charakteristikách – pohlaví, věku, výšce a informaci o kouření pomocí jednorozměrné i vícerozměrné lineární regrese. Výsledný model pro přirozený logaritmus FEV je založen na věku, výšce a kouření, přičemž s vyšším věkem a výškou mají děti lepší dýchací funkce (větší lnFEV), zatímco kouření zhoršuje lnFEV, přestože z jednorozměrného pohledu by se mohlo zdát, že kouření dýchací funkce zlepšuje. Pro tento model byla provedena analýza reziduí a hledání odlehlých a vlivných pozorování. Nebylo zjištěno žádné závažné porušení předpokladů modelu, nebyla zaznamenána ani příliš vlivná pozorování, která by po jejich odstranění z analýzy změnila výsledky finálního modelu.