Analýza procentuálního zastoupení tělesného tuku v závislosti na různých faktorech

Tereza Fucsiková

08/09/2021

Úvod do datasetu

Data, která máme pro zkoumání k dispozici, byla poskytnuta Dr. A. Garth Fisherem, prvním ředitelem Centra pro výzkum lidské výkonnosti (Human Performance Research Center) v Utahu. Dataset obsahuje procentuální zastoupení tělesného tuku ve dvou odlišných měření (podle Siri a Brozka), věk, hmotnost, výšku a dalších 10 naměřených obvodů různých částí těla (např. obvod břicha, stehna,...) u 252 mužů.

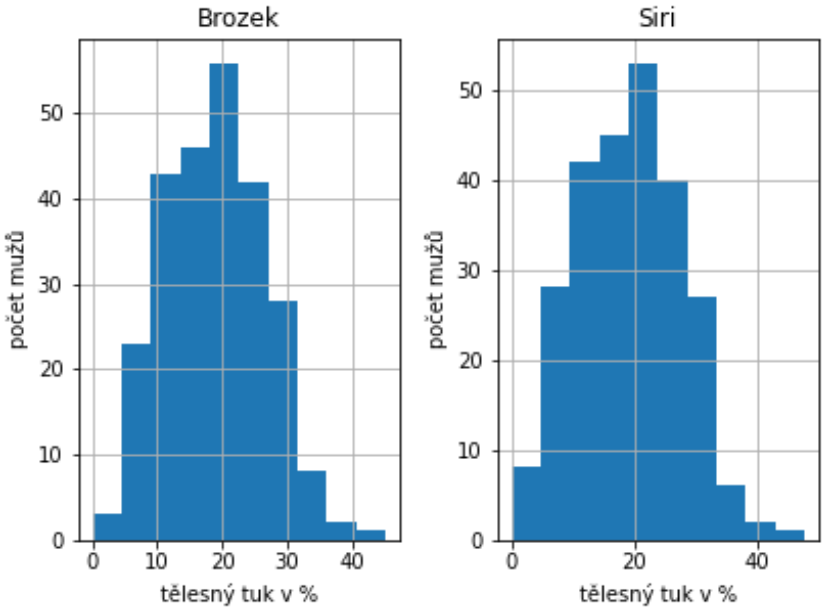
Cílem našeho zkoumání bude porovnání těchto dvou různých rovnic měření tělesného tuku, obojí v závislosti na ostatních faktorech, jako je výše zmíněný obvod břicha a jiné. Vliv faktorů na množství tuku v těle je dalším předmětem zkoumání.

Grafická deskriptivní analýza

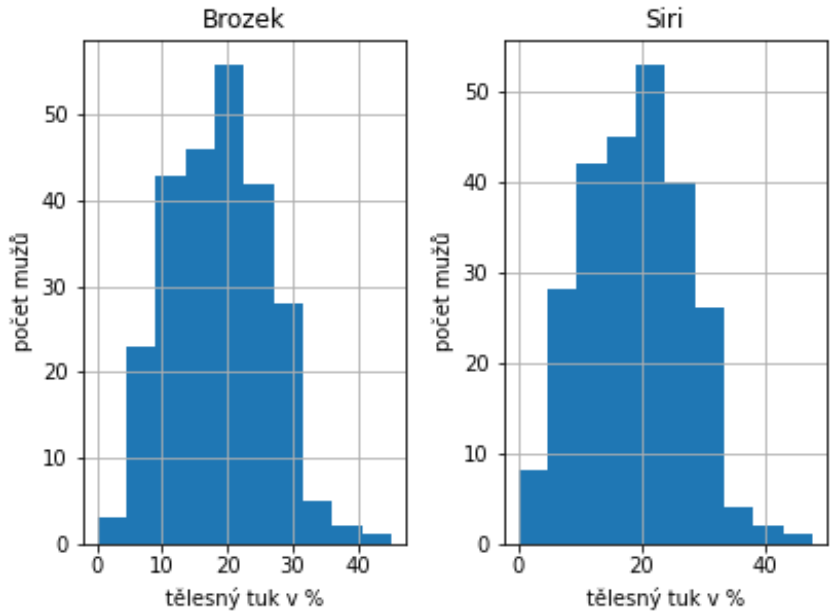
Nejprve si data vykreslíme pomocí histogramů a boxplotů, čímž získáme ucelenou podobu dat, pomocí které lze provést případnou korekci odlehlých pozorování. Korekcí se v tomto případě myslí vyřazení dat, která se odchylují od ostatních hodnot. Těmito daty mohou být osamocené sloupce v histogramech, či kružnice zobrazované mimo boxploty.

Celkem nám dataset poskytuje 16 faktorů ovlivňujících množství tuku v těle. Pro přehlednost si z nich vybereme 9 faktorů, se kterými budeme nadále pracovat. Jmenovitě to jsou: věk, váha, výška, hmotnost bez tuku a obvody krku, břicha, boků, stehna a bicepsu.

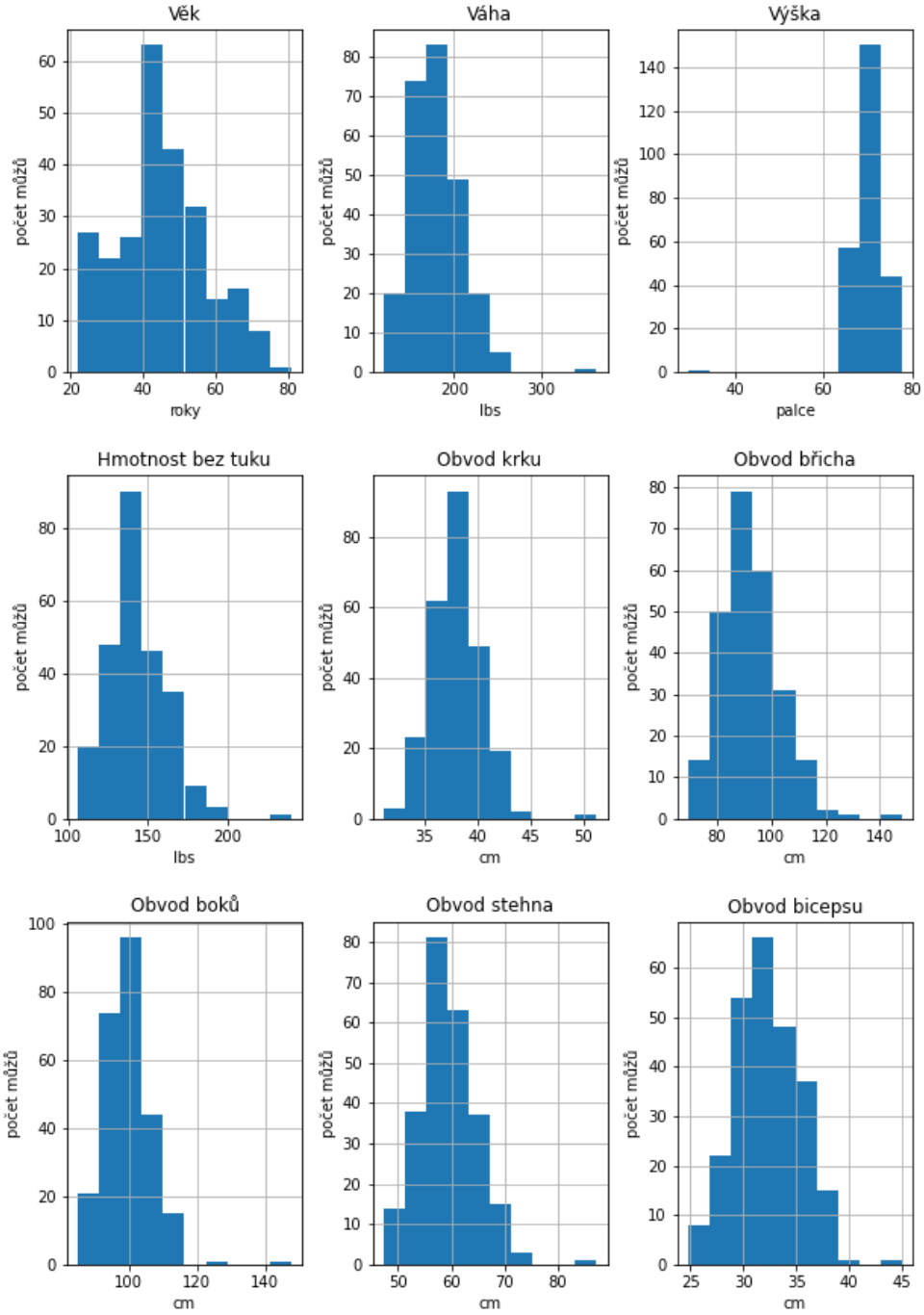
Histogramy



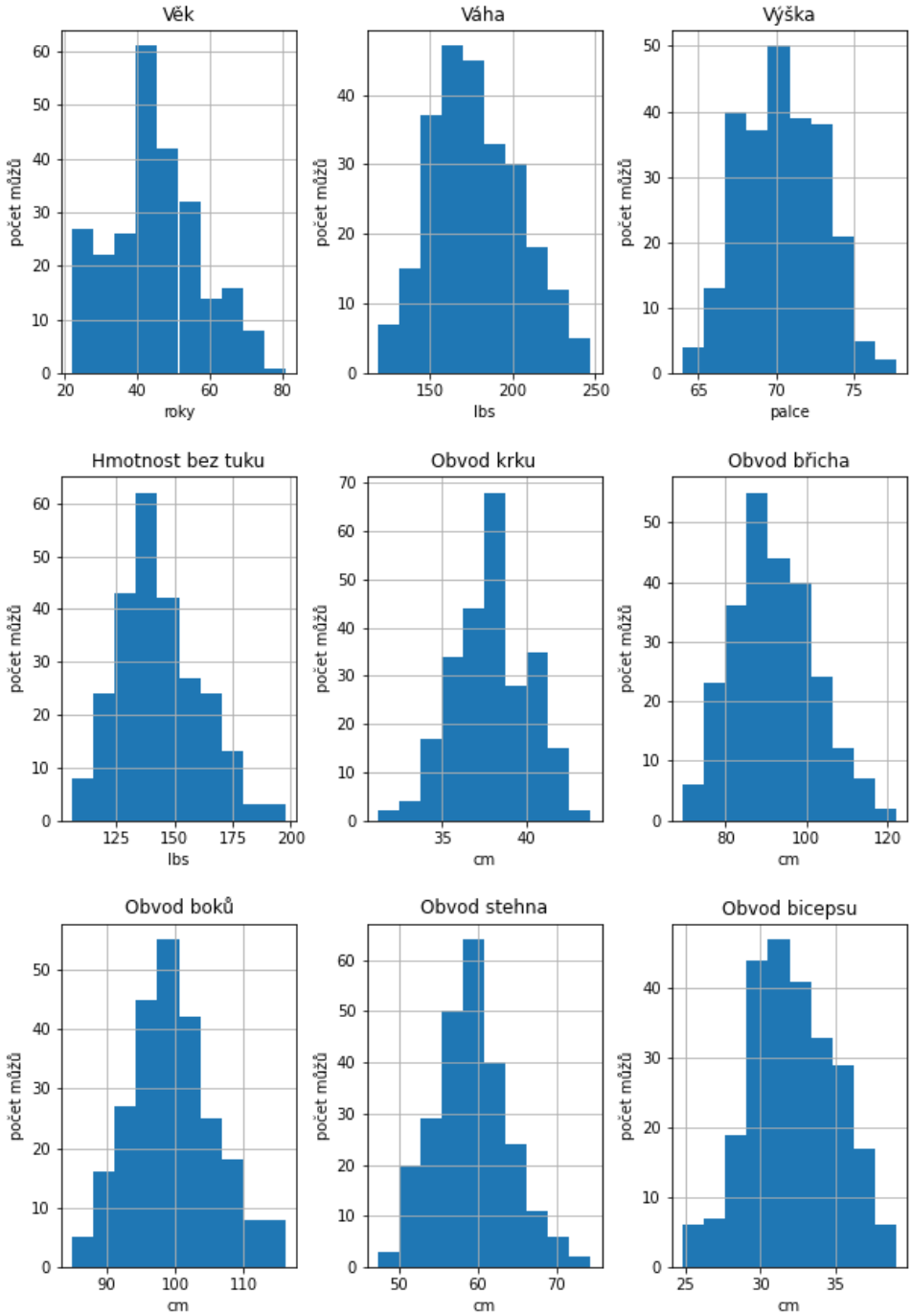
Obr. 1: Histogramy měření tělesného tuku podle Brozka a Siri v %



Obr. 2: Histogramy měření tělesného tuku podle Brozka a Siri v % po korekci

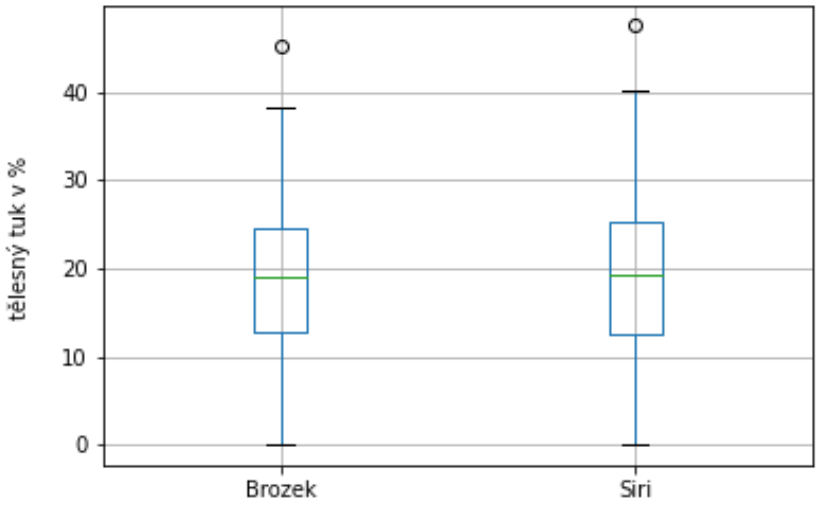


Obr. 3: Histogramy faktorů ovlivňujících zastoupení tuku v těle

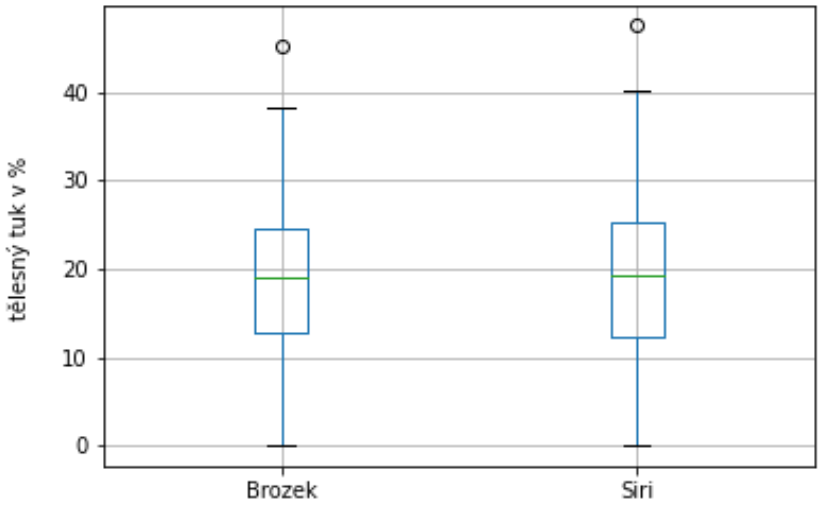


Obr. 4: Histogramy faktorů ovlivňujících zastoupení tuku v těle po korekci

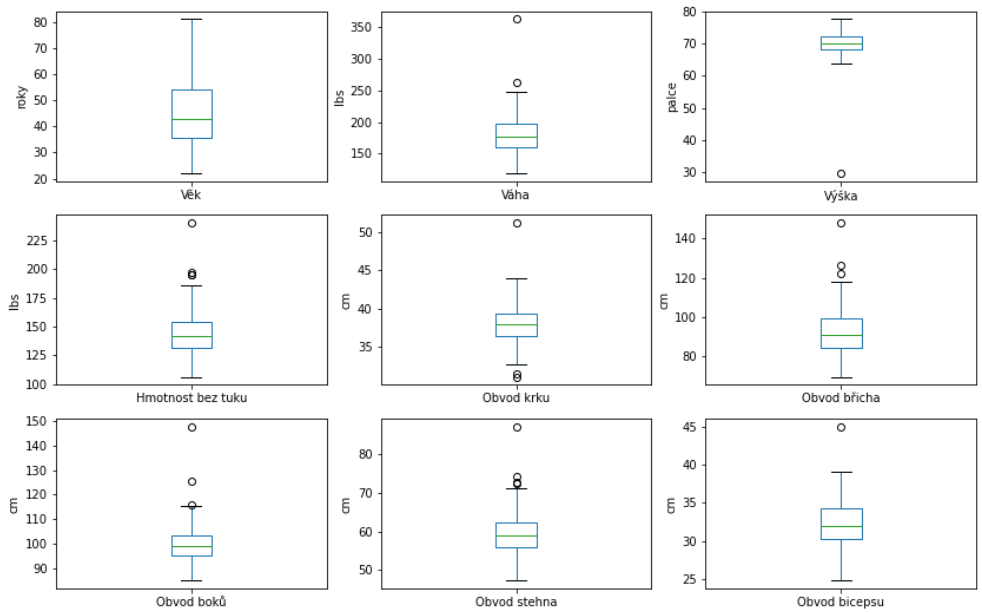
Boxploty



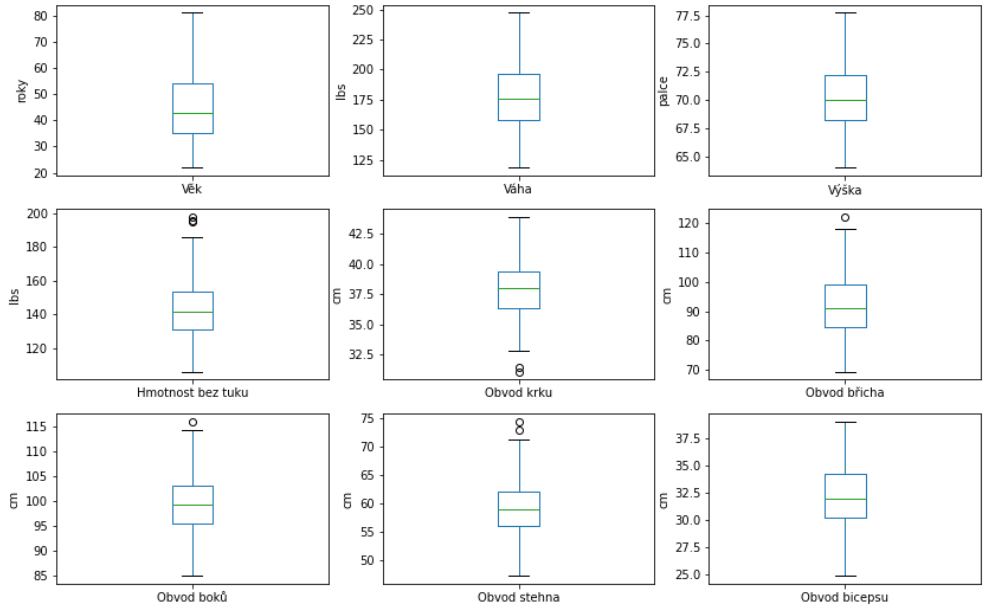
Obr. 5: Boxploty měření tělesného tuku podle Brozka a Siri v %



Obr. 6: Boxploty měření tělesného tuku podle Brozka a Siri v % po korekci



Obr. 7: Boxploty faktorů ovlivňujících zastoupení tuku v těle



Obr. 8: Boxploty faktorů ovlivňujících zastoupení tuku v těle po korekci

Korekce zmiňovaná v grafech proběhla odstraněním tří pozorovaných osob: pozorování č. 39, 41 a 42.

Numerická deskriptivní analýza

V této části si popíšeme studované proměnné. U každé proměnné určíme následující hodnoty:

$\boldsymbol{\mu}$ ..... střední hodnota
s ..... směrodatná odchylka
min ..... minimum
max ..... maximum
$\boldsymbol{x_{1/4}}$ ..... kvantil pro 1/4
$\boldsymbol{x_{1/2}}$ ..... medián
$\boldsymbol{x_{3/4}}$ ..... kvantil pro 3/4

	$\boldsymbol{\mu}$	**s**	**min**	**max**	$\boldsymbol{x_{1/4}}$	$\boldsymbol{x_{1/2}}$	$\boldsymbol{x_{3/4}}$
Brozek (%)	18.938492	7.750856	0	45.1	12.8	19	24.6
Siri (%)	19.150794	8.368740	0	47.5	12.475	19.2	25.3
Věk (roky)	44.884921	12.602040	22	81	35.75	43	54
Váha (lbs)	178.924405	29.389160	118.5	363.15	159	176.5	197
Výška (palce)	70.148810	3.662856	29.5	77.75	68.25	70	72.25
Hmotnost bez tuku(lbs)	143.713889	18.231642	105.9	240.5	131.35	141.55	153.875
Obvod krku (cm)	37.992063	2.430913	31.1	51.2	36.4	38	39.425
Obvod břicha (cm)	92.555952	10.783077	69.4	148.1	84.575	90.95	99.325
Obvod boků (cm)	99.904762	7.164058	85	147.7	95.5	99.3	103.525
Obvod stehna (cm)	59.405952	5.249952	47.2	87.3	56	59	62.35
Obvod bicepsu (cm)	32.273413	3.021274	24.8	45	30.2	32.05	34.325

Po korekci vypadá tabulka následovně:

	$\boldsymbol{\mu}$	**s**	**min**	**max**	$\boldsymbol{x_{1/4}}$	$\boldsymbol{x_{1/2}}$	$\boldsymbol{x_{3/4}}$
Brozek (%)	18.770683	7.643186	0	45.1	12.8	19	24.5
Siri (%)	18.969478	8.252221	0	47.5	12.4	19.2	25.2
Věk (roky)	44.883534	12.677708	22	81	35	43	54
Váha (lbs)	177.743173	26.548892	118.5	247.25	158.25	176	196.75
Výška (palce)	70.309237	2.620052	64	77.75	68.25	70	72.25
Hmotnost bez tuku(lbs)	143.210843	17.15177	105.9	197.7	131.2	141.4	153.8
Obvod krku (cm)	37.923695	2.270578	31.1	43.9	36.4	38	39.4
Obvod břicha (cm)	92.150602	9.997795	69.4	122.1	84.5	90.9	99.1
Obvod boků (cm)	99.546988	6.241943	85	116.1	95.5	99.3	103.1
Obvod stehna (cm)	59.196386	4.849462	47.2	74.4	56	58.9	62.1
Obvod bicepsu (cm)	32.200402	2.916216	24.8	39.1	30.2	32	34.3

Vlastní analýza

Pro naše účely budeme sestavovat dva vícerozměrné lineární regresní modely, které budou zkoumat procentuální zastoupení tělesného tuku (dle Brozka nebo Siri) v závislosti na ostatních devíti zvolených faktorech. Abychom zajistili funkční model, je nutné ověřit následující předpoklady:

1. **Množství dat je vetší, než počet zkoumaných parametrů** (tj. v našem případě má být počet mužů podstupujících měření větší, než jednotlivé kategorie měření).
   - Tento předpoklad je automaticky splněn, v obou modelech máme k dispozici 249 dat a 9 kategorií měření.

2. **Matice tvořená vysvětlujícími proměnnými** (tj. rozměru 249 x 9) **má lineárně nezávislé sloupce**, nebo-li plnou hodnost.
   - Porovnáváme hodnoty R^2-statistiky s významností vysvětlujících proměnných, jež zjišťujeme pomocí t-testu. Předpoklad není splněn, pokud dostaneme velkou hodnotu R^2-statistiky, ale skoro žádná proměnná nebude významná.

3. **Náhodné chyby mají normální rozdělení a stejný rozptyl $\boldsymbol{\sigma^2}$**.
   - Ke splnění tohoto předpokladu nám pomůže analýza reziduí modelu. Rezidua jsou vlastně odhady náhodných chyb, tudíž by pro ně měly být splněny stejné předpoklady. Ověření splnění předpokladů zjistíme z grafů (případně lze použít i testy normality rozdělení).

Sestavení modelu pro rovnici Siri/Brozka

Jelikož se nám modely odlišují pouze vysvětlovanou proměnnou, sestavíme jeden společný model, v němž budeme nadále diskutovat obě varianty zvlášť. Uvažujeme tedy model

$$Y_i^{(k)} = \beta_0 + \sum_{j=1}^{9} \beta_j x_{ij} + e_i \ , \quad i = 1, ..., 249, \ \ k = \{1, 2\},$$

kde $Y_i^{(1)}$ je vektor měření tuku podle rovnice Brozka, $Y_i^{(2)}$ poté podle rovnice Siri (tzv. vysvětlované proměnné). Vysvětlujícími proměnnými x_{ij} jsou naše faktory ovlivňující množství tuku v těle. Dále označují β_0, β_j neznámé regresní parametry a e_i představuje vektor náhodných chyb.

Model pro rovnici Brozka nám poskytl následující výsledky:

vysvětlující proměnné	**odhad parametrů β_j**	**standardní chyba**	**p-hodnota**
Konstanta ($\boldsymbol{\beta_0}$)	-12.7771	5.261	0.016
Věk ($\boldsymbol{x_1}$)	0.0131	0.008	0.121
Váha ($\boldsymbol{x_2}$)	0.3418	0.017	0
Výška ($\boldsymbol{x_3}$)	0.2425	0.051	0
Hmotnost bez tuku($\boldsymbol{x_4}$)	-0.518	0.011	0
Obvod krku ($\boldsymbol{x_5}$)	0.106	0.066	0.107
Obvod břicha ($\boldsymbol{x_6}$)	0.1012	0.029	0.001
Obvod boků ($\boldsymbol{x_7}$)	0.0241	0.043	0.574
Obvod stehna ($\boldsymbol{x_8}$)	0.1341	0.04	0.001
Obvod bicepsu ($\boldsymbol{x_9}$)	0.1133	0.048	0.019

Odstraněním statisticky nevýznamných veličin, tedy veličin jejichž p-hodnota je větší než uvažovaná hladina významnosti $\alpha = 0.05$, dostaneme finální model. V našem případě odstraňujeme veličiny x_1, x_5, x_7, tedy věk, obvod krku a obvod boků.

vysvětlující proměnné	**odhad parametrů β_j**	**standardní chyba**	**p-hodnota**
Konstanta ($\boldsymbol{\beta_0}$)	-7.473	4.223	0.078
Váha ($\boldsymbol{x_2}$)	0.3475	0.016	0
Výška ($\boldsymbol{x_3}$)	0.2241	0.05	0
Hmotnost bez tuku($\boldsymbol{x_4}$)	-0.5142	0.011	0
Obvod břicha ($\boldsymbol{x_6}$)	0.1252	0.027	0
Obvod stehna ($\boldsymbol{x_8}$)	0.1108	0.033	0.001
Obvod bicepsu ($\boldsymbol{x_9}$)	0.1323	0.047	0.005

Výsledný model je tvaru

$$Y_i^1 = -7.473 + 0.3475x_{i2} + 0.2241x_{i3} - 0.5142x_{i4} + 0.1252x_{i6} + 0.1108x_{i8} + 0.1323x_{i9}, \quad i = 1, ..., 249,$$

s mírou kvality $R^2 = 0.974$, která se blíží jedné a zajišťuje nám tak, že náš model velmi dobře popisuje data. Nyní můžeme tvrdit, že 2. předpoklad je splněn. Máme pouze tři statisticky nevýznamné veličiny z devíti a model funguje velmi dobře. Standardní chyba je 4.224.

Model pro rovnici Siri má výsledky:

vysvětlující proměnné	**odhad parametrů β_j**	**standardní chyba**	**p-hodnota**
Konstanta ($\boldsymbol{\beta_0}$)	-15.2818	5.745	0.008
Věk ($\boldsymbol{x_1}$)	0.0155	0.009	0.092
Váha ($\boldsymbol{x_2}$)	0.3678	0.0197	0
Výška ($\boldsymbol{x_3}$)	0.2617	0.055	0
Hmotnost bez tuku($\boldsymbol{x_4}$)	-0.5586	0.012	0
Obvod krku ($\boldsymbol{x_5}$)	0.1063	0.072	0.139
Obvod břicha ($\boldsymbol{x_6}$)	0.1094	0.032	0.001
Obvod boků ($\boldsymbol{x_7}$)	0.0334	0.047	0.476
Obvod stehna ($\boldsymbol{x_8}$)	0.1354	0.044	0.002
Obvod bicepsu ($\boldsymbol{x_9}$)	0.1343	0.052	0.011

Odstraněním stejných statisticky nevýznamných veličin získáváme hodnoty:

vysvětlující proměnné	**odhad parametrů β_j**	**standardní chyba**	**p-hodnota**
Konstanta ($\boldsymbol{\beta_0}$)	-9.3958	4.612	0.043
Váha ($\boldsymbol{x_2}$)	0.3742	0.018	0
Výška ($\boldsymbol{x_3}$)	0.2427	0.055	0
Hmotnost bez tuku($\boldsymbol{x_4}$)	-0.5548	0.012	0
Obvod břicha ($\boldsymbol{x_6}$)	0.1378	0.029	0
Obvod stehna ($\boldsymbol{x_8}$)	0.1115	0.036	0.002
Obvod bicepsu ($\boldsymbol{x_9}$)	0.1533	0.052	0.003

Výsledný model je tvaru

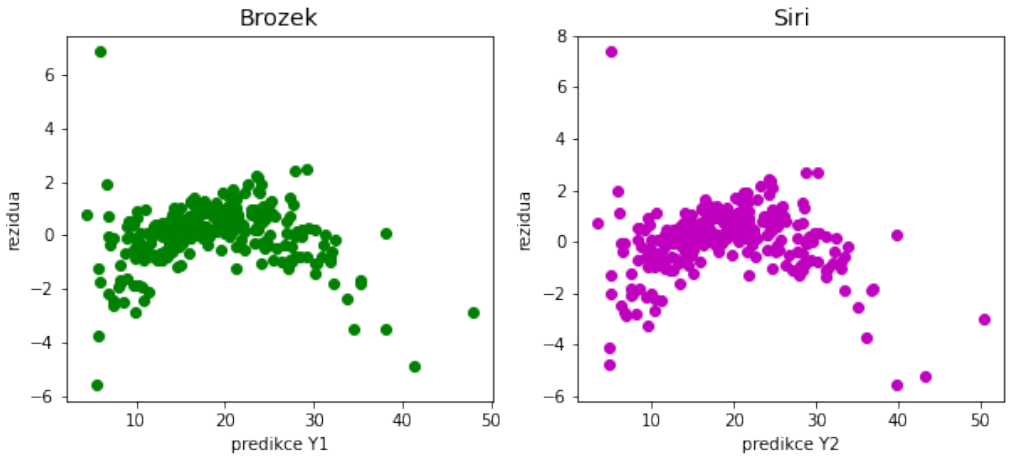
$$Y_i^1 = -9.3958 + 0.3742x_{i2} + 0.2427x_{i3} - 0.5548x_{i4} + 0.1378x_{i6} + 0.1115x_{i8} + 0.1533x_{i9}, \quad i = 1, ..., 249,$$

s mírou kvality $R^2 = 0.973$ a 2. předpoklad je opět splněn. Standardní chyba je 4.613.

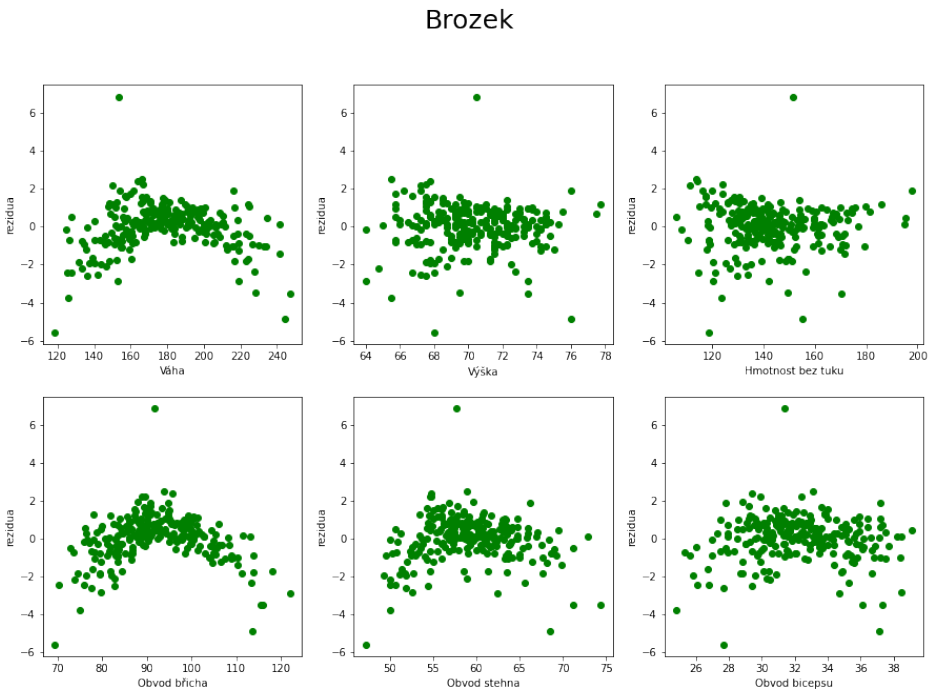
Analýza reziduí grafickými nástroji

Nyní si zobrazíme 3 různé grafy reziduí obou modelů, pomocí kterých budeme moci diskutovat splnění normality, tedy našeho posledního předpokladu. V prvních dvou typech grafu se zaměříme na to, zda jsou rezidua rovnoměrně rozptýlena. Jakýkoliv trend v grafech může znamenat porušení podmínky normality. Na posledním grafu, tzv. kvantilovém grafu reziduí, bychom měli vidět data kopírující přímku.

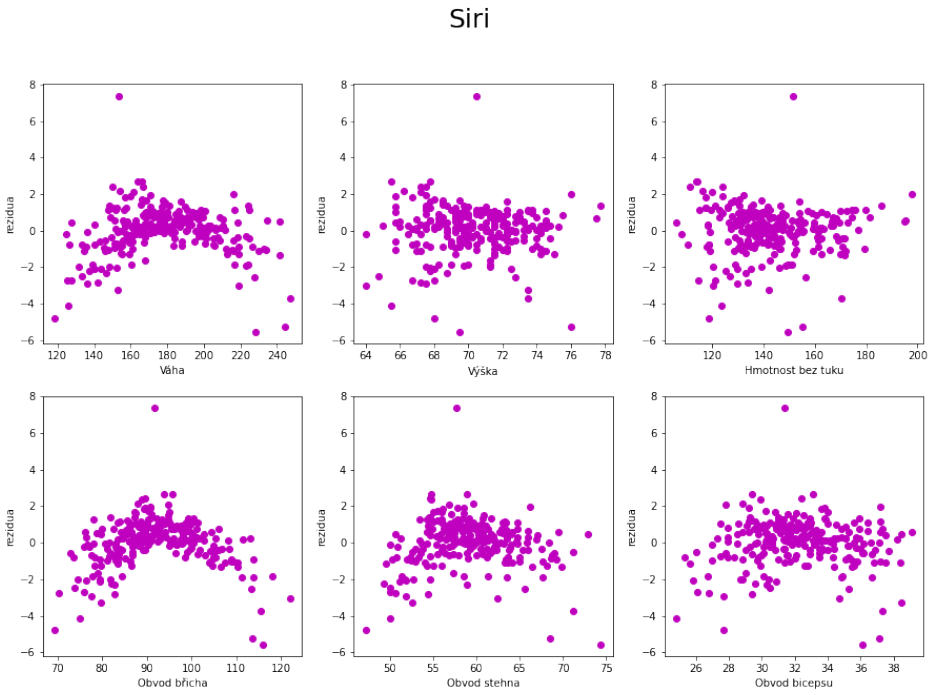
Na následujících grafech si můžeme povšimnout trendu v podobě paraboly. Normalita tedy není jednoznačně splněna. V případě závislosti na konkrétních faktorech, lze podle grafů nalézt rovnoměrné rozmístění, parabolický trend je však významný především u váhy a obvodu břicha.



Obr. 9: Graf reziduí v závislosti na predikovaných hodnotách

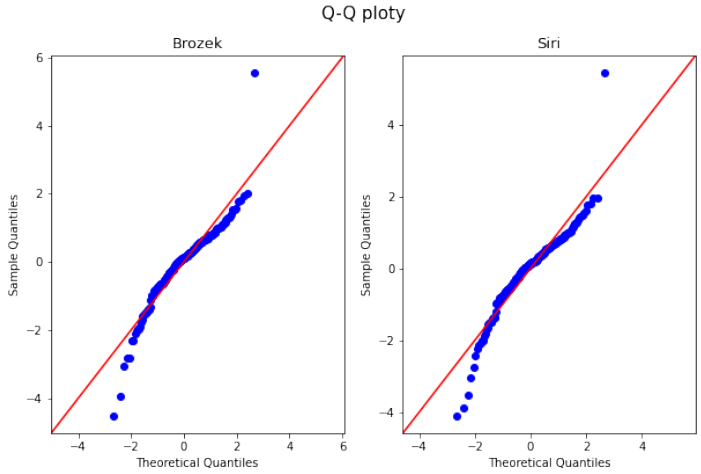


Obr. 10: Graf reziduí v závislosti vysvětlujících proměnných pro rovnici Brozka



Obr. 11: Graf reziduí v závislosti vysvětlujících proměnných pro rovnici Siri

U kvantilových grafů vyvrací normalitu hodnoty na chvostech, které nám správně nekopírují přímku. Zároveň si můžeme povšimnout, že se nám jedna hodnota nachází velmi daleko od těch ostatních.



Obr. 12: Kvantiové grafy reziduí

Závěr

Zkoumali jsme procentuální zastoupení tělesného tuku v závislosti na různých faktorech. K dispozici jsme měli dvě rovnice, které množství tuku v těle měří.

Grafická část deskriptivní statisktiky nám ukázala odlehlá měření, která jsme poté vyřadili, abychom dosáhli lepších výsledků.

Následně jsme sestavili dva lineární regresní modely, zvlášť pro rovnici Brozka a Siri. Důležité bylo splnění předpokladů, kde ovšem nastaly komplikace v případě normality. V grafech jsme si mohli povšimnout tzv. trendů, které byly v rozporu s požadovaným rovnoměrným rozdělením.

Z obou modelů jsme se dozvěděli, že věk, obvod krku a obvod boků nemají vliv na množství tuku v těle. Získali jsme vysokou míru kvality modelů, která činila u rovnice Brozka 0.974 a u rovnice Siri 0.973. Rovnice Siri měla zároveň i větší standardní chybu, z čehož vyplývá, že rovnice Brozka by měla býti o něco přesnější. Obě rovnice se však jevily celé naše zkoumání jako velmi podobné, což bylo patrné jak z grafů, tak i z velmi podobných výsledných hodnot zkoumání.