Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra matematiky

Diplomová práce

Metody identifikace odlehlých pozorování

Plzeň, 2015 Bc. Zuzana Loudová

Oficiální zadání práce

Prohlášení

Prohlašuji, že jsem diplomovou práci na téma „Metody identifikace odlehlých pozorování“ vypracovala samostatně a výhradně za použití citovaných pramenů.

V Plzni dne 11. 5. 2015

……

Bc. Zuzana Loudová

Poděkování

Tímto bych chtěla poděkovat vedoucí mé diplomové práce RNDr. Blance Šedivé, Ph.D za cenné rady poskytnuté při zpracování tématu a také za propůjčené materiály.

Abstrakt

Diplomová práce „Metody identifikace odlehlých pozorování“ se zabývá detekcí odlehlých hodnot v datových souborech. Tento proces je velmi důležitý před dalším zpracováním dat, aby nedocházelo ke zkresleným výsledkům. Cílem této práce je tak uvést metody vhodné pro detekci odlehlých pozorování v datových souborech jednorozměrných i vícerozměrných. V úvahu budou brány také časové řady a detekce odlehlých pozorování v regresi. Popsány budou metody parametrické i neparametrické. Pro zvolené metody pak budou provedeny numerické experimenty, díky kterým budou odhaleny výhody i nevýhody různých postupů. V závěru diplomové práce budou zvolena reálná data, ve kterých budou odlehlá pozorování hledaná pomocí uvedených metod.

Součástí diplomové práce jsou kódy provedené v programu Matlab obsahující veškeré výpočty.

Klíčová slova: odlehlé hodnoty, jednorozměrné metody, vícerozměrné metody, odlehlé hodnoty v regresi, numerické experimenty, metody založené na vzdálenosti bodů, metody založené na hustotě bodů, Mahalanobisova vzdálenost.

Abstract

The diploma thesis „Methods of identification of outliers” deals with a detection of outliers in sets of data. To prevent the distortion of results, it is very important to carry out this process before further data processing. The aim of this thesis is to introduce methods for detection of outliers in univariate and multivariate sets of data. Time series and outliers in regression will be taken into consideration, too. The parametric and nonparametric methods will be described. According to the selected methods, numerical experiments will be carried out which will reveal the advantages and disadvantages of different kinds of methods. In conclusion of the thesis, the real data for detection of outliers will be chosen and different kinds of methods will be used.

This diploma thesis also includes the codes performed in Matlab, which contain all the calculations.

Keywords: outliers, univariate methods, multivariate methods, outliers in regression, numerical experiments, distance-based methods, density-based methods, Mahalanobis distance.

Obsah

1 Úvod

V současné době, kdy je stále nutné zpracovávat velké množství dat a získávat tak důležité informace, je zároveň velmi důležité, aby právě tato data byla co nejpřesnější a hlavně správná, bez chyb způsobených lidským faktorem nebo například poruchou stroje. Jsou-li tato data posbíraná např. samotným počítačem nebo lidským elementem, je tak nejprve nutné prověřit kvalitu těchto dat.

Jedním z důvodů prověřování kvality dat je možná přítomnost tzv. odlehlých pozorování. A právě odlehlá pozorování a metody jejich identifikace jsou tématem této diplomové práce. Detekce odlehlých pozorování je velmi žádaná v mnoha oblastech, ať už se jedná o lékařství, kde se mohou hledat odlehlé hodnoty v reakcích na léčbu, přes různé výrobní procesy, kde mohou odlehlé hodnoty udávat zmetky bez možnosti prodeje, až po pojišťovnictví, kde odlehlé hodnoty upozorňují na možné pojistné podvody. Zkrátka, odlehlé hodnoty se mohou vyskytovat ve většině datových souborů a je velmi užitečné tyto hodnoty odhalit a následně s nimi pracovat podle uvážení. V některých případech je vyloučit z dalšího analyzování dat, aby nedocházelo ke zkreslování statistických charakteristik celého souboru, v jiných případech dále pracovat právě s nimi, jako například při identifikaci neobvyklého chování pojištěnce.

Tato diplomová práce tak slouží k uvedení několika metod detekce odlehlých pozorování, ať už v jednorozměrných datových souborech nebo vícerozměrných. Dále bude popsána také detekce odlehlých hodnot v regresi a pro všechny zmíněné metody budou provedeny numerické experimenty pro lepší pochopení jejich fungování. Tyto numerické experimenty budou provedeny v programu Matlab pomocí uměle vygenerovaných souborů dat.

Na závěr budou vhodné metody využity také při detekci odlehlých pozorování v reálných datech.

2 Teorie

Pojem odlehlá pozorování, neboli často uváděné v angličtině „outliers“, lze definovat několika způsoby. Barnett, Lewis [1] definují odlehlá pozorování jako *an observation (or subset of observation) which appears to be inconsistent with the remander of that set of data.* Pod pojmem odlehlé hodnoty (outliers) tedy budou v celé práci uvažovány takové hodnoty, které významně vybočují ze zkoumaného souboru a tím ovlivňují jeho vlastnosti pro další případné statistické analyzování souboru.

Avšak je velmi složité přesně definovat pojem „významně vybočující ze souboru“. Je totiž dosti složité identifikovat, zda se jedná o odlehlé pozorování nebo pouze o extrémní hodnotu, která však nese důležitou informaci pro další analýzu. Je tedy klíčové správně rozlišit, zda je zkoumaná hodnota nedílnou součástí souboru nebo je naopak chybná.

Při sběru dat mohou nastat situace, kdy dochází k náhodným chybám, ale také k hrubým chybám, které jsou zapříčiněny lidským pochybením (ať už úmyslným nebo nechtěným) nebo případným chybným nastavením měřidel určených ke sběru dat. Ovšem vždy je na pozorovateli, zda podezřelou hodnotu z odlehlosti vyhodnotí jako chybnou nebo pouze extrémní. Je totiž v nejvyšším zájmu pozorovatele uchovat všechny důležité informace pro další statistické analýzy dat. Odhalí-li však zcela jistě chybnou hodnotu, je vhodné ji z dalšího analyzování vyřadit. Jistě však může označit jako chybnou hodnotu pouze tu, která neodpovídá předpokladům celého souboru dat. Pokud například pozorovatel analyzuje teplotní údaje v našich podnebných podmínkách, pouhým okem vyhodnotí jako odlehlou hodnotu například 216°C. Přitom mohlo dojít pouze k nechtěné lidské chybě, kdy bylo opomenuto zaznamenat desetinnou čárku. Hodnota 21,6°C by totiž byla odpovídající. Nemá-li však pozorovatel možnost zkonzultovat své domněnky s kompetentní osobou, která byla přítomna sběru dat a která by tak mohla případné chyby opravit, je nejvhodnější tyto hodnoty vyloučit z dalších výpočtů, aby nedocházelo ke zkreslování vlastností celého souboru dat.

A právě k identifikaci takovýchto hrubých chyb slouží několik metod, které budou zmíněny níže.

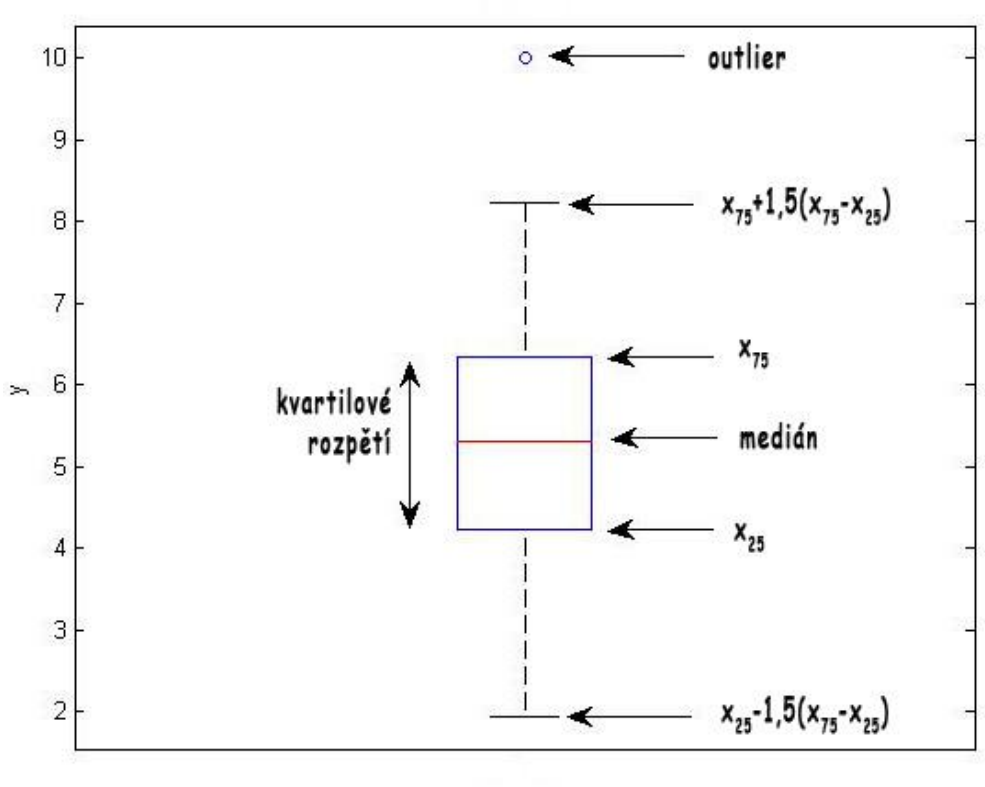
2.1 Grafické metody

Tyto metody jsou těmi nejintuitivnějšími metodami. Jsou velmi názorné a dochází-li k prezentaci výsledků, jsou velmi žádané.

2.1.1 Boxplot

Boxplot, neboli krabicový diagram, patří mezi grafické metody využívající kvartilů datového souboru. Střední část diagramu je zdola ohraničena 1. kvartilem (x_{25}) a shora pak 3. kvartilem (x_{75}) a informuje pozorovatele také o mediánu, variabilitě za kvartily či odlehlých hodnotách, které mohou být znázorněny jako samostatné body. Boxploty jsou grafickou, ale také neparametrickou metodou, nevyžadují tedy, aby zkoumaný soubor dat odpovídal předem známému rozdělení [2,4].

Boxploty mohou být znázorněny mnoha způsoby. Mohou být vertikální či horizontální. Vždy budou obsahovat krabicovou část, avšak znázorněné linie z ní vycházející jsou různé. Mohou zahrnovat minimum a maximum zkoumaných dat nebo jsou například omezené 1,5 násobkem kvartilového rozpětí. Dolní hodnota linie je tak dána jako $x_{25} - 1{,}5(x_{75} - x_{25})$ a horní hodnota je dána jako $x_{75} + 1{,}5(x_{75} - x_{25})$. Odlehlé hodnoty jsou pak znázorněny samostatnými body, jak již bylo zmíněno výše. Pro lepší názornost bude zobrazen boxplot s liniemi o velikosti 1,5 násobku kvartilového rozpětí.



Obrázek 2.1.1: Boxplot

2.1.2 Bagplot

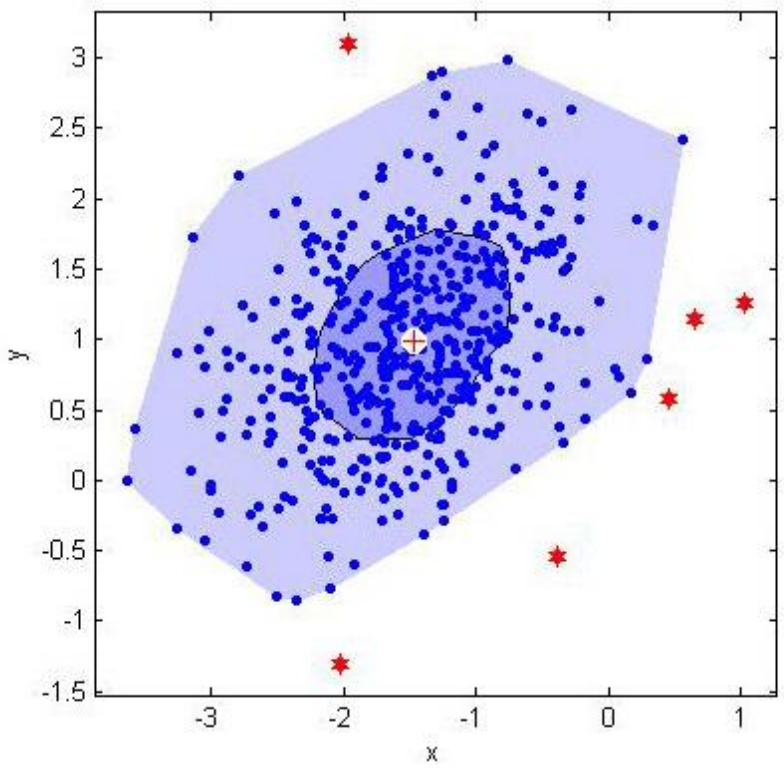
Bagplot je dalším grafickým zobrazením možných odlehlých hodnot. V podstatě je bagplot dvourozměrnou verzí boxplotu, kterým je možné detekovat outliers jednorozměrných dat. Dle [18] jsou hlavními částmi bagplotu takzvaný „bag“, který obsahuje 50 % veškerých dat datového souboru, dále tzv. plot („fence“), který separuje data od outliers a poslední částí je tzv. smyčka („loop“) znázorňující body mimo bag, ale uvnitř plotu.

Bagplot obsahuje také medián, který je v dvourozměrné verzi bagplotu umístěn v centru souboru dat, neboli nejhlouběji v datovém souboru. Je tak analogií mediánu boxplotu v dvourozměrném prostoru. Smyčka je pak analogií linií v boxplotu vycházející z krabicové části.

Pomocí bagplotu tak lze identifikovat některé charakteristiky dat, jako je například dvourozměrný medián, rozpětí (velikost bagu), korelaci (orientaci bagu) či pro tuto práci stěžejní outliers, které se nachází mimo smyčku.

Konstrukce bagplotu je založena na tzv. hloubce bodu $\theta \in \mathbb{R}^2$ v poloprostoru, označována jako $ldepth(\theta, Z)$, kde $Z = (z_1, z_2, \dots, z_n)$ jsou pozorování souboru dat a platí $z_i = (x_i, y_i)$. Tento pojem představil John Tukey, podle něhož se označuje jako Tukeyho hloubka bodu. Jedná se o nejmenší počet pozorování, která mohou být obsažené v jakémkoliv uzavřeném poloprostoru obsahující zvolený bod. Pojem „oblast hloubky“ D_k je pak definován jako všechny body θ, jejichž hloubka $ldepth(\theta, Z) \geq k$. Oblasti hloubky jsou konvexní mnohoúhelníky a platí $D_{k-1} \subset D_k$. Medián vícerozměrných dat je tak definován jako bod θ s největší hloubkou. Bag je pak tvořen za podmínky $\#D_k \leq \frac{n}{2} < \#D_{k-1}$, kde $\#D_k, \#D_{k-1}$ označují počet pozorování v jednotlivých oblastech. Plot je získán „zvětšením“ bagu nejčastěji faktorem 3. Hodnoty za plotem jsou hodnoceny jako outliers.

Pro grafické znázornění bagplotu je využit matlabovský kód ze zdroje [28], jehož konstrukce je založena právě na výše popsaném principu hloubky bodů. Bagplot je znázorněn na následujícím *Obrázku 2.1.2*, kde je s datovým souborem zobrazen také Tukeyho medián (červený křížek v bílém kolečku), bag (plocha kolem Tukeyho mediánu tmavší modré barvy), smyčka (plocha kolem bagu světlejší modré barvy) a nakonec také outliers, které jsou označeny jako červené hvězdy.



Obrázek 2.1.2: Bagplot

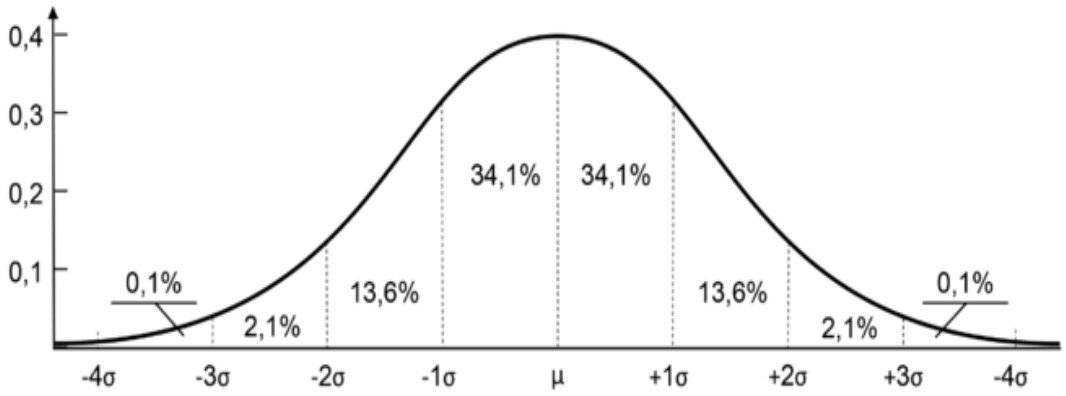
2.2 Jednorozměrné metody

Tyto metody identifikace odlehlých pozorování patří mezi ty nejjednodušší a také nejstarší metody. Dále je pak můžeme rozdělit na parametrické, které předpokládají vztah souboru dat s nějakým předem známým rozdělením, a na neparametrické, u kterých takovýto vztah nepředpokládáme.

Parametrické metody jsou tedy využívány, je-li předem známé statistické rozdělení a jeho parametry. Nejčastěji se v praxi objevují data, která odpovídají normálnímu rozdělení, a právě pro taková data existuje více metod identifikace outliers.

2.2.1 Pravidlo 3-sigma

Pravidlo 3-sigma je pravidlo, které v případě souboru dat pocházejícího z normálního rozdělení říká, že téměř všechny hodnoty souboru by se měly nacházet do vzdálenosti 3 směrodatných odchylek od hodnoty průměru.



Obrázek 2.2.1: Znázornění pravidla 3-sigma (Zdroj [3])

Pravidlo tedy udává, že v případě normálního rozdělení by se ve vzdálenosti 1 směrodatné odchylky od průměru mělo nacházet 68,27 % hodnot, ve vzdálenosti 2 směrodatných odchylek pak 95,45 % hodnot a ve vzdálenosti 3 směrodatných odchylek 99,73 % hodnot zkoumaného souboru. Hodnoty, které se nachází ve větší vzdálenosti od průměru než právě 3 směrodatné odchylky, považujeme za možné odlehlé hodnoty.

2.2.2 Grubbsův test

Dalším parametrickým testem předpokládající normální rozdělení je dle [1] Grubbsův test. Tento test se obvykle využívá při větším počtu dat zkoumaného souboru. Tato data je nejprve nutné vzestupně uspořádat následujícím způsobem $x_1 \leq x_2 \leq \cdots \leq x_n$.

Toto uspořádání slouží ke zjištění odlehlosti nejnižší hodnoty (x_1) nebo nejvyšší hodnoty (x_n) pomocí testovacích kritérií

$$T_1 = \frac{\bar{x} - x_1}{s} \qquad \text{nebo} \qquad T_n = \frac{x_n - \bar{x}}{s} \tag{1}$$

kde $\bar{x}$ je chápáno jako aritmetický průměr počítaný ze všech hodnot souboru $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ pak jako výběrová směrodatná odchylka.

Vypočtené statistiky T_1 a T_n je pak nutné porovnat s kritickou hodnotou T_α pro požadovanou hladinu významnosti α. Je-li hodnota $T_1 > T_\alpha$ či $T_n > T_\alpha$, pak je hodnoceno toto pozorování jako odlehlé.

Výpočet kritických hodnot Grubbsova testu je dán následovně [13]

$$T_\alpha = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/n,n-2}}{n-2+t^2_{\alpha/n,n-2}}} \tag{2}$$

kde $t^2_{\alpha/n,n-2}$, značí kvadrát kvantilu studentova rozdělení se stupněm volnosti $n-2$ a hladinou významnosti α/n. Tabulku kritických hodnot jednotlivých počtů pozorování je možné nalézt v *Příloze 1*.

2.2.3 Dean-Dixonův test

Další parametrickou metodou odhalení odlehlých pozorování založenou na předpokladu normálního rozdělení je Dean-Dixonův test známý též jako Q-test. Tento test je však určený spíše pro menší soubory dat. Obvykle se využívá při 3 - 10 pozorovaných hodnotách [14]. Opět je nutné hodnoty vzestupně uspořádat následujícím způsobem $x_1 \leq x_2 \leq \cdots \leq x_n$. Dále je možné zkoumat, zda je minimální hodnota x_1 nebo maximální hodnota x_n potenciálně odlehlou. Pro testování těchto hodnot slouží testové statistiky

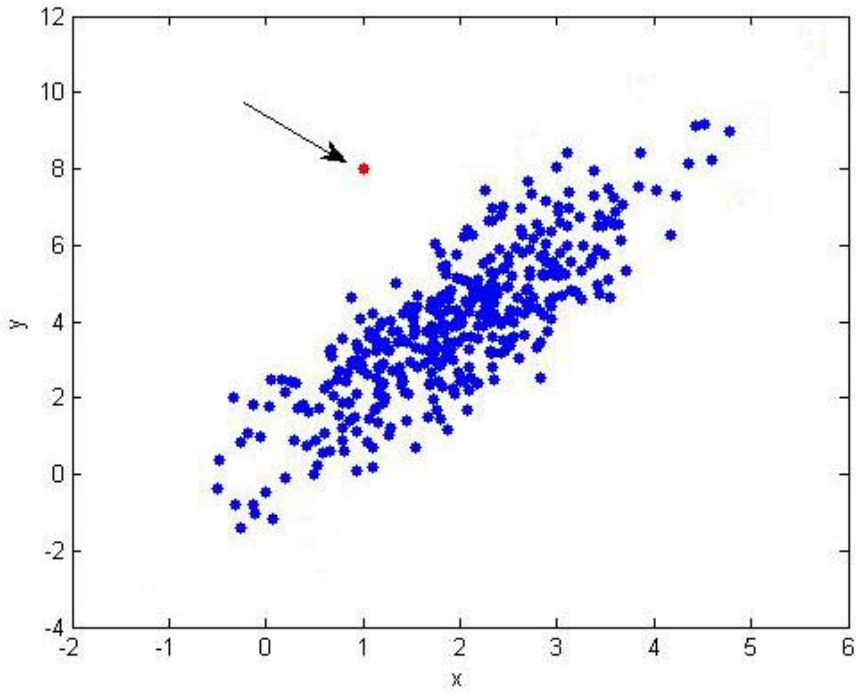
$$Q_1 = \frac{x_2 - x_1}{x_n - x_1} \qquad \text{nebo} \qquad Q_n = \frac{x_n - x_{n-1}}{x_n - x_1} \tag{3}$$

kde jmenovatel $x_n - x_1$ lze také nazývat variační rozpětí. Aby bylo možné rozhodnout, zda je minimální, případně maximální, hodnota souboru odlehlým pozorováním je nutné vypočtenou testovou statistiku porovnat s kritickou hodnotou Q_α. Platí-li $Q_1 > Q_\alpha$ či $Q_n > Q_\alpha$, je možné s pravděpodobností $1-\alpha$ říci, že zkoumaná hodnota je odlehlá od zbytku souboru. Tabulku kritických hodnot jednotlivých počtů pozorování je možné nalézt v *Příloze 2*.

Výše zmíněný Grubbsův a Dean-Dixonův test jsou schopny odhalit v jednom kroku pouze jednu odlehlou hodnotu maximální, případně minimální. Aby bylo jisté, že v souboru již jiná taková hodnota není, je nutné provádět test opětovně, do doby, kdy soubor již žádnou odlehlou hodnotu nedetekuje.

2.3 Mnohorozměrné metody

V mnoha případech mnohorozměrného pozorování nelze detekovat odlehlé hodnoty, je-li každá proměnná posuzována nezávisle na druhé, případně ostatních. Odlehlé hodnoty je tak nutné posuzovat pomocí mnohorozměrných metod, které zohledňují interakci mezi jednotlivými proměnnými. Jednoduchým příkladem je možné znázornit, jaký problém by mohl nastat, pokud by nebyly využity mnohorozměrné metody.



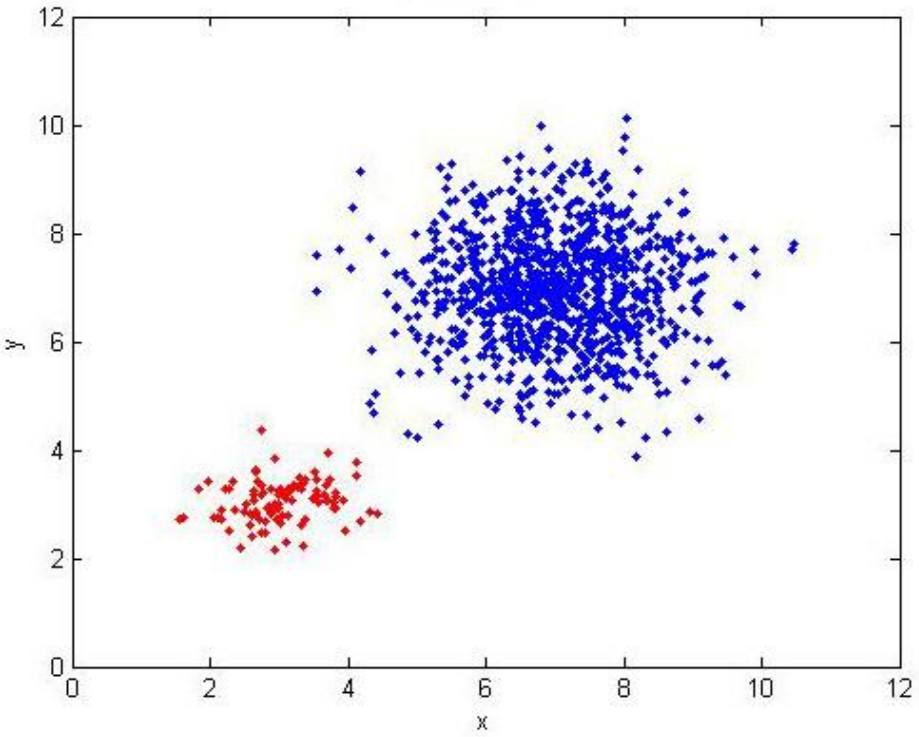
Obrázek 2.3.1: Odlehlá hodnota mnohorozměrných dat

Červeně označený bod v *Obrázku 2.3.1* lze považovat za odlehlou hodnotu mnohorozměrných dat. Pokud by však tato hodnota byla posuzována pro každou proměnnou zvlášť, nebyla by odlehlou ani v jednom případě, tedy ve směru osy x ani ve směru osy y. A právě z tohoto důvodu je bezpodmínečně nutné testovat mnohorozměrná data pomocí metod pro ně určených.

V případě mnohorozměrných dat se mohou dle [15] vyskytovat dva efekty, a to tzv. *masking effect* a *swamping effect.*

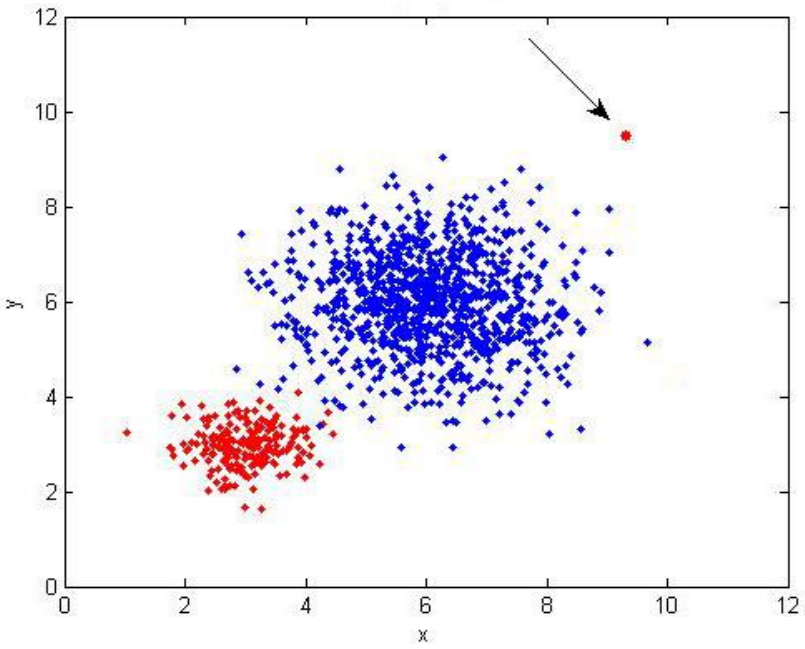
- *Masking effect (maskování)* – tento efekt znamená, že outlier je těžce detekovaný z důvodu blízkosti dalších pozorování. Jinými slovy si lze představit shluk pozorování, přičemž k souboru dat patří ještě další pozorování, která jsou však vzdálena tomuto shluku a tvoří vlastní menší shluk. Jedno pozorování z menšího shluku je odlehlé vzhledem k většině dat, avšak není odlehlé vzhledem k ostatním hodnotám menšího shluku. Tyto

hodnoty však výrazně vychylují odhady parametrů, čímž se ovšem také mění parametry definice odlehlosti hodnot (neboli tyto hodnoty nemusejí být odhaleny jako odlehlé). Na následujícím *Obrázku 2.3.2* jsou znázorněny možné odlehlé hodnoty červeně. Kterákoliv červená hodnota tak může být posuzována jako odlehlá od shluku modrých, zatímco v rámci červených pozorování odlehlá není.



Obrázek 2.3.2: Masking effect

- *Swamping effect (přeplnění, zaplavení)* – tento efekt znamená, že správná hodnota byla detekována jako odlehlá v důsledku jiného pozorování. Efekt nastává v případě, pokud menší skupina odlehlých pozorování (v *Obrázku 2.3.3* zobrazených červeně) opět zkreslila odhady parametrů a tím se z původně správné hodnoty nacházející se poblíž hlavní skupiny dat stala hodnota odlehlá. V obrázku je tato hodnota opět zabarvená červeně, avšak pokud by se v souboru dat nenacházel menší shluk hodnot, nebyla by hodnocena jako odlehlá.



Obrázek 2.3.3: Swamping effect

Mnohorozměrné metody detekce outliers mohou být rozděleny na statistické metody, které jsou založené na odhadnutých parametrech rozdělení a metody podobné data-mining metodám, které jsou nezávislé na parametrech.

2.3.1 Mahalanobisova vzdálenost

Mahalanobisova vzdálenost je dobře známým kritériem, které závisí na odhadnutých parametrech mnohorozměrného rozdělení. Je dáno n pozorování z p-dimenzionálního souboru dat, který lze popsat pomocí vektoru výběrových průměrů $\bar{\boldsymbol{x}}_n$ a výběrovou kovarianční maticí $\boldsymbol{\Sigma}_n$, kde

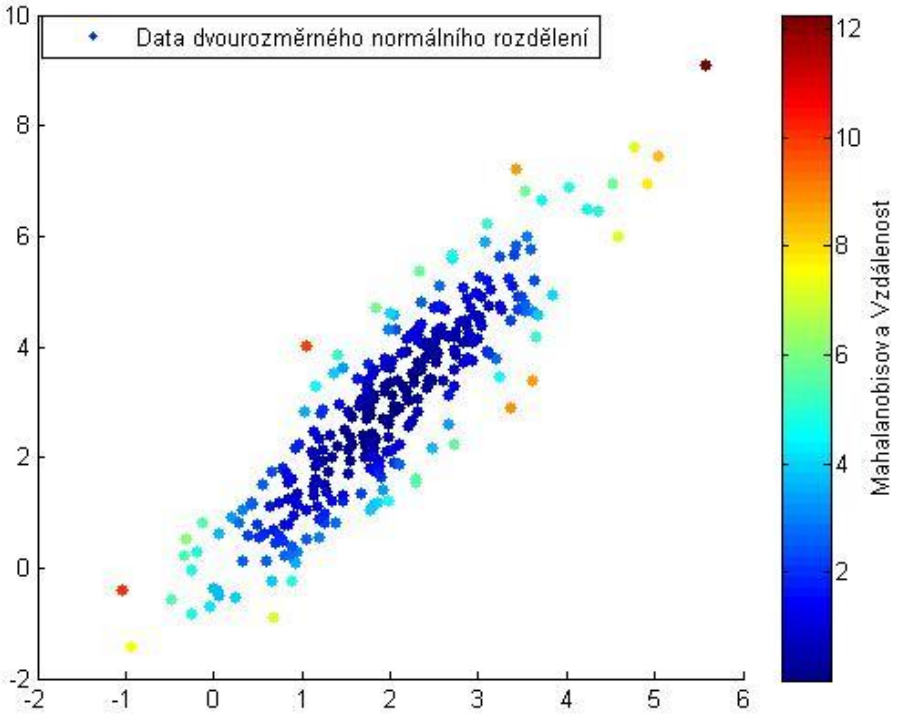
$$\boldsymbol{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)^T \tag{4}$$

Mahalanobisova vzdálenost je pak dle [15] definována pro každé pozorování $i, i = 1, \dots, n$ jako

$$d_{M(i)} = \sqrt{\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)^T {\boldsymbol{\Sigma}_n}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)} \tag{5}$$

Pozorování s vysokou hodnotou Mahalanobisovy vzdálenosti jsou pak podezřelá z odlehlosti. Je nutné brát v úvahu, že v adekvátnosti použití Mahalanobisovy vzdálenosti jako kritéria posuzování odlehlých pozorování hrají významnou roli masking efekt i swamping efekt. Konkrétně masking

efekt může snižovat Mahalanobisovu vzdálenost outliers právě zkreslením odhadu průměrů. A na druhou stranu swamping efekt může navyšovat vzdálenost neodlehlých hodnot.



Obrázek 2.3.4: Znázornění Mahalanobisovy vzdálenosti vícerozměrných normálně rozdělených dat

V *Obrázku 2.3.4* jsou zobrazena náhodně generovaná data dvourozměrného normálního rozdělení a pomocí barevné škály označena pozorování podezřelá z odlehlosti, přičemž těmito pozorováními jsou červené až černé body s nejvyšší Mahalanobisovou vzdáleností.

2.3.2 Distance-based metody

Tyto neparametrické metody byly dle [8] poprvé zmíněny dvojicí Knorr a Ng a jsou založeny na principu nejbližších sousedů. Předpokládá se tedy, že větší vzdálenost od nejbližšího souseda detekuje možné outliers. Metod založených na zkoumání vzdáleností mezi jednotlivými pozorováními je velká řada a v této práci budou zmíněny jen některé.

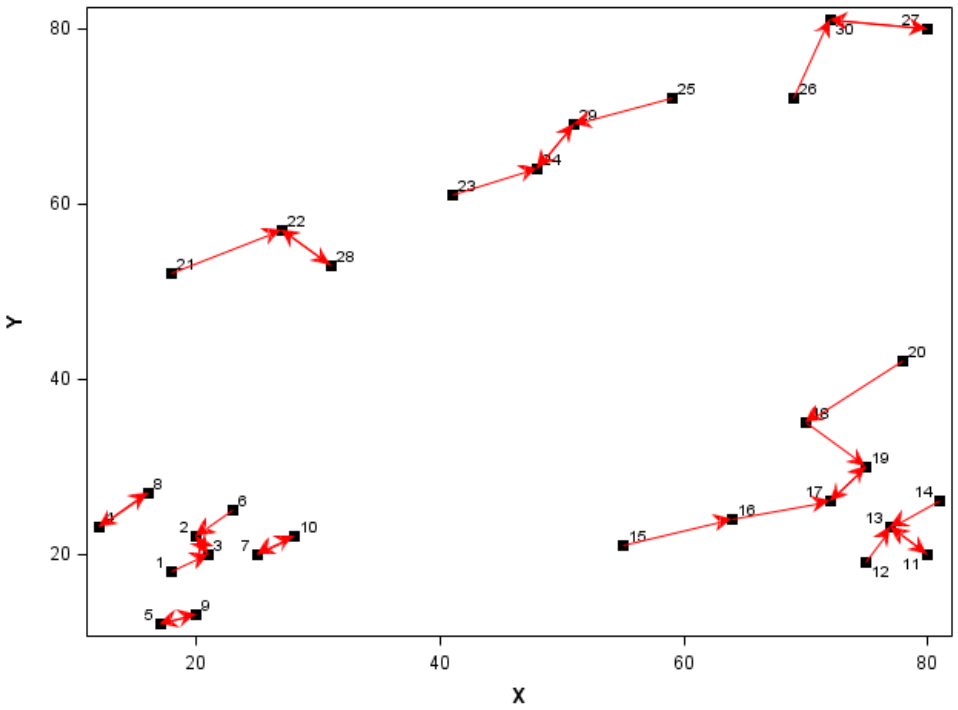
Dle [9] existuje několik definic odlehlých pozorování z pohledu distance-based metod.

- Odlehlé hodnoty jsou taková pozorování s méně než *k* sousedy v souboru dat, kde soused je pozorování, které se nachází ve vzdálenosti *R*. Zvolit však apriori pevně okolí *R* je velmi složité.
- Odlehlými hodnotami je *n* pozorování s největší vzdáleností od *k*-tého nejbližšího souseda (*kNN* definice).

2.3.2.1 Metody k-nejbližších sousedů

Metody *k*-nejbližších sousedů, známé také jako *kNN* algoritmy, jsou neparametrické metody užívané pro klasifikaci dat, kdy se zjišťuje příslušnost určitého pozorování k dané třídě, či v regresi, kdy může být požadováno rozdělení dat dle určité vlastnosti do skupin a následně pokračovat v analýzách.

K zobrazení slouží také tzv. *kNN grafy*, což jsou grafy, které spojují bod *p* s jeho nejbližším sousedem *q*. Často jsou tyto grafy zobrazovány jako neorientované či orientované a zároveň nesymetrické. Je-li *q* nejbližším sousedem *p,* není nutně bod *p* nejbližším sousedem bodu *q.* Tvorba tohoto grafu však není zcela triviální [10].



Obrázek 2.3.5: kNN graf nejbližšího souseda (Zdroj [29])

V *Obrázku 2.3.5* lze vidět již zmiňovanou nesymetričnost. Zatímco například bod 21 má jako nejbližšího souseda bod 22, ten nemá jako nejbližšího souseda opět bod 21, ale bod 28.

2.3.2.2 Nested loop algoritmus

Hlavní myšlenkou tohoto algoritmu je, že pro každý bod souboru bude sledováno jeho *k*-nejbližších sousedů. Následně algoritmus zjišťuje vzdálenosti od ostatních bodů a nalezne-li takový bod, který má vzdálenost menší než jiný z původních *k*-nejbližších sousedů, nahradí tento bod. Odlehlé hodnoty jsou pak ty s nejvyšší vzdáleností od zvoleného bodu. Tento algoritmus však není příliš vhodný pro analyzování rozsáhlých souborů dat, z důvodu jeho náročnosti [8].

2.3.2.3 KDIST a MeanDIST

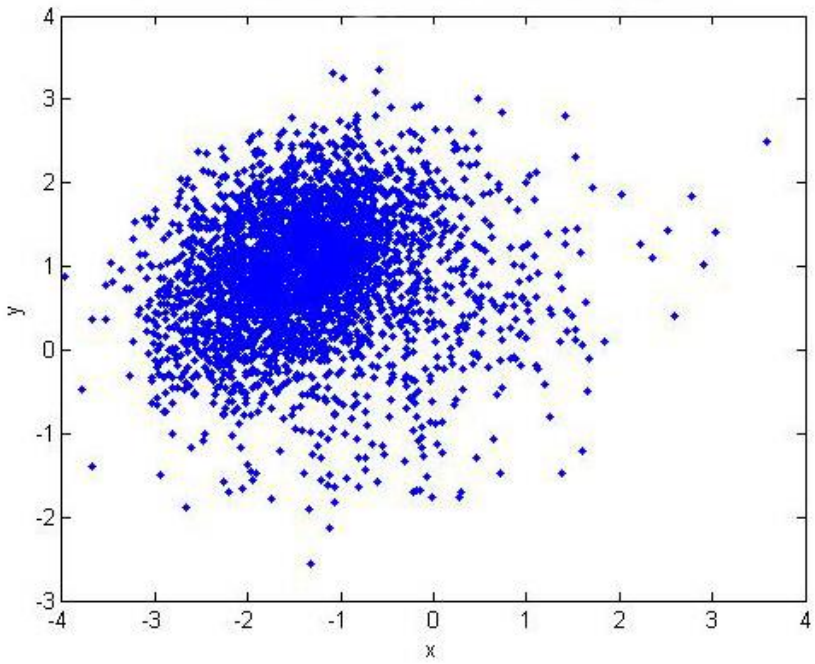
Dalšími metodami založenými na principu *k*-nejbližších sousedů jsou metody *KDIST* a *MeanDIST* [10]. Jsou opět založeny na výpočtu vzdáleností všech bodů a jejich *k*-nejbližších sousedů. Tyto metody se liší v dalším použití těchto vzdáleností. Využije-li se průměr těchto vzdáleností, jedná se o metodu *MeanDIST,* zatímco maximum těchto vzdáleností indikuje metodu *KDIST.* Každému pozorování je tak přiřazena jedna hodnota dle zvolené metody. Seřazení pozorování pak umožní výpočet diferencí přiřazených hodnot sousedních bodů. Pomocí těchto diferencí lze pak identifikovat odlehlé hodnoty a to díky stanovené hranici. Překročí-li velikost diference tuto hranici, pozorování je hodnoceno jako odlehlé. Tato hranice je dána

$$T = max(L_i - L_{i-1}) \cdot t \tag{6}$$

kde L_i je *KDIST* nebo *MeanDIST i*-tého vektoru a $t \in [0,1]$ je pozorovatelem stanovený parametr.

2.3.3 Density-based metody

Mnohorozměrná data je možné zkoumat z hlediska vzdáleností mezi sebou, jak bylo zmíněno výše, ale také z hlediska hustoty bodů v okolí daného pozorování. Základem těchto metod je tedy idea, že outliers jsou body s menší hustotou sousedů. Tyto metody je tedy možné využít právě ve chvíli, kdy po zobrazení pozorování je patrný rozdíl hustot bodů v jednotlivých oblastech definovaného prostoru. Na následujícím *Obrázku 2.3.6* je tak vidět, že data v levé části grafu mají vyšší hustotu sousedů. Možné odlehlé hodnoty se tak budou nacházet v části pravé, kde je hustota sousedů pro jednotlivé body nižší.



Obrázek 2.3.6: Graf zobrazující hustotu souboru dat

2.3.3.1 Local outlier faktor

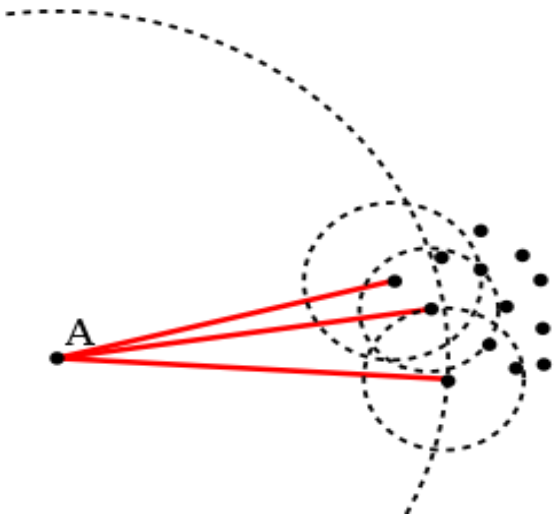
Metodou identifikace outliers založené na výpočtu hustot kolem bodů je metoda *local outlier factor (LOF).* Česky lze tuto metodu vyjádřit jako Místní míru odlehlosti [11].

Označení $k\_distance(A)$ je chápáno jako vzdálenost pozorování *A* od jeho *k*-nejbližšího souseda. Dále je definována množina $N_k(A)$, což je množina *k*-nejbližších sousedů. Využívá se také tzv. dosažitelné vzdálenosti (reachability distance). Zdroj [11] pak udává

$$reachability\_distance_k(A,B) = max\{k\_distance(B), d(A,B)\} \tag{7}$$

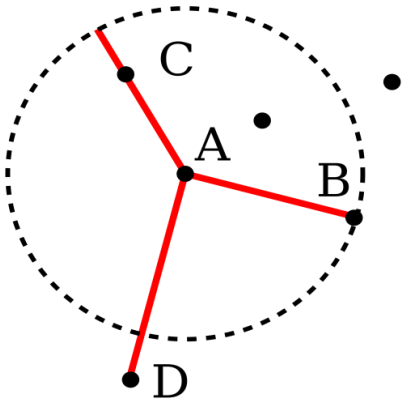
Pojmem dosažitelné vzdálenosti pozorování A od pozorování B je tedy chápána skutečná vzdálenost těchto dvou bodů, přinejmenším však *k_distance(B).* Pak objekty náležící ke *k*-nejbližším sousedům bodu B jsou považovány za stejně vzdálené. Je však nutné mít stále na paměti, že toto není vzdálenost v matematickém smyslu slova, neboť není symetrická.

Tyto pojmy je možné demonstrovat na následujícím *Obrázku 2.3.7*, kde je znázorněn bod A a jeho *k*-nejbližší sousedé. V tomto konkrétním případě je pak $k = 3$. Je tady patrné, že bod A má menší hustotu než jeho sousedé.



Obrázek 2.3.7: Porovnání hustoty bodu A s hustotami jeho sousedů (Zdroj [11])

Dosažitelnou vzdálenost je pak možné demonstrovat na dalším *Obrázku 2.3.8*. Ten zobrazuje, že body B a C mají stejnou dosažitelnou vzdálenost $(k = 3)$, zatímco D není *k*-nejbližší soused bodu A, a tak bude $reachability\_distance_k(D,A)$ rovna jeho skutečné vzdálenosti od bodu A.



Obrázek 2.3.8: Dosažitelná vzdálenost (Zdroj[11])

Pomocí *Obrázku 2.3.8* je také lépe představitelný pojem dosažitelné vzdálenosti. Ta je dána např. pro body A, C jako $reachability\_distance_k(A,C) = max\{k\_distance(C), d(A,C)\}$. Je-li parametr $k = 3$, pak i bod C patří mezi k-nejbližších sousedů bodu A, a tak je dosažitelná vzdálenost rovna vyšší hodnotě, což je konkrétně v tomto případě $k\_distance(C)$. Zatímco dosažitelná vzdálenost mezi body A a D je brána jako jejich pravá vzdálenost $d(A,D)$.

Dále lze dle [11] definovat tzv. místní dosažitelnou hustotu (local reachability density) pozorování A jako

$$lrd(A) = 1 \Big/ \left(\frac{\sum_{B \in N_k(A)} reachability\_distance_k(A,B)}{|N_k(A)|} \right) \tag{8}$$

Dá se také interpretovat jako převrácená hodnota průměrné dosažitelné vzdálenosti bodu A od jeho sousedů. Jsou-li nalezeny duplicitní hodnoty, může $lrd(A)$ nabývat také nekonečna.

Místní faktor odlehlosti je pak dle [12] definován jako

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} \tag{9}$$

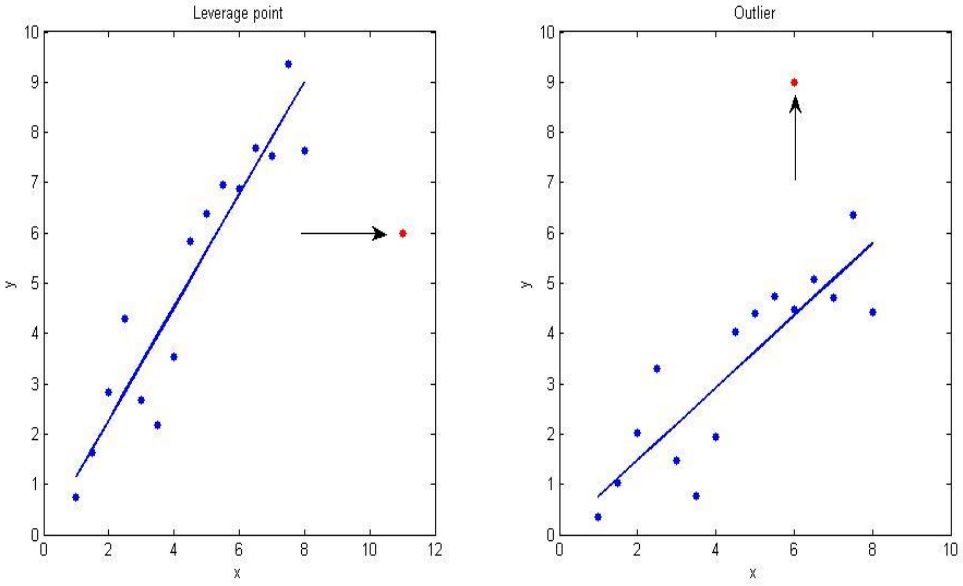
Tento faktor lze také interpretovat jako průměr poměru místní dosažitelné hustoty sousedů a místní dosažitelné hustoty bodu A. Vypočtená hodnota pak udává odlehlost bodu. Pokud je tato hodnota blízká 1, pak lze říci, že místní dosažitelné hustoty jsou srovnatelné a pozorování tak není odlehlé. Zatímco převyšuje-li *LOF* hodnotu 1, detekuje outlier. Hodnota převyšující 1 je způsobena vyšší hustotou sousedů a naopak nižší hustotou bodu A.

2.4 Outliers v regresi

Regresní analýza je velmi využívaným nástrojem při zkoumání datových souborů v praxi, neboť na základě získaných informací z dat umožňuje pozorovateli tvořit predikci budoucího vývoje. Právě predikce je nejžádanějším výsledkem různých statistických analýz, a je tedy požadováno, aby byla co možná nejpřesnější. A právě možné odlehlé hodnoty v datech mohou způsobit volbu špatného modelu nebo „jen“ špatný odhad budoucích hodnot.

Proto je velmi důležité před samotnou regresní analýzou posoudit kvalitu dat a to zejména s ohledem na možné vlivné body, což jsou takové body, které mohou významně ovlivnit regresní analýzu. Takovými body mohou být opět správné hodnoty, které ovšem vybočují od ostatních. Takové body označujeme jako *body s vysokým vlivem*, které je však nutné ponechat v datovém souboru, neboť v sobě ukrývají důležité informace vývoje dat. Opět však může docházet také k tzv. *hrubým chybám*, ať už chybou lidského faktoru nebo špatným nastavením měřidel.

Speciálně v regresi jsou pak pod pojmem *outliers* chápána odlehlá pozorování ve směru vysvětlované proměnné Y, zatímco odlehlá pozorování ve směru vysvětlující proměnné X budou označována *leverage points* [19].



Obrázek 2.4.1: Zobrazení Outliers a Leverage points

V regresi se k identifikaci vlivných bodů používají míry detekce založené na reziduích. Přičemž existuje několik typů reziduí, která jsou využívána.

2.4.1 Typy reziduí

V dalším textu bude uvažován lineární normální regresní model $Y \sim N(X\beta, \sigma^2 I)$, pomocí kterého budou definovány typy reziduí.

- **Klasická rezidua**

Pro lineární normální regresní model $Y \sim N(X\beta, \sigma^2 I)$ označme $e = Y - X\beta$ rezidua vyjadřující rozdíl mezi naměřenými hodnotami $Y = (y_1, y_2, \dots y_n)^T$ a vyrovnanými hodnotami $\hat{Y} = X^T b$ odhadnutými metodou nejmenších čtverců, kde $b = (X^T X)^{-1} X^T Y$.

Klasická rezidua lze tedy vyjádřit jako

$$e_i = Y_i - \hat{\hat{Y}}_i \tag{10}$$

Tato rezidua mají nulovou střední hodnotu, nekonstantní rozptyl a jsou navzájem závislá.

- **Normovaná rezidua**

Normovaná rezidua lze vyjádřit

$$e_i^{(N)} = \frac{e_i}{s} \tag{11}$$

tedy poměrem klasických reziduí a odmocněným odhadem reziduálního rozptylu, který je dán jako $s^2 = \frac{e^T e}{n-p}$, kde n je počet pozorování a p je počet sloupců matice X.

Tato rezidua mají opět nulovou střední hodnotu, ale jejich rozptyl je $var\left(e_i^{(N)}\right) = (1 - h_{ii})$, kde h_{ii} jsou diagonální prvky projekční matice H, která je dána jako $H = X(X^T X)^{-1} X^T$. Normovaná rezidua jsou navzájem závislá.

- **Studentizovaná rezidua (jackknife)**

Studentizovaná rezidua lze vyjádřit

$$e_i^{(T)} = \frac{e_i}{s_{(-i)}\sqrt{1 - h_{ii}}} \tag{12}$$

kde $s_{(-i)}^2$ je odhad reziduálního rozptylu bez *i* tého pozorování

$$s_{(-i)}^2 = \frac{\left(Y - Xb_{(-i)}\right)^T \left(Y - Xb_{(-i)}\right)}{n - p - 1} \tag{13}$$

Přičemž výpočetní tvar $s^2_{(-i)}$ je založen na odhadu reziduálního rozptylu a projekční matici H [20, 21]

$$s^2_{(-i)} = \frac{e^T e - \frac{e_i^2}{1-h_{ii}}}{n-p-1} = \frac{(n-p)s^2 - \frac{e_i^2}{1-h_{ii}}}{n-p-1} \tag{14}$$

V *rovnici (13)* je využit výraz $b_{(-i)}$, což je odhad parametrů modelu β získaný na základě všech n pozorování bez i-tého, tedy

$$b_{(-i)} = \left(X^T_{(-i)} X_{(-i)}\right)^{-1} X^T_{(-i)} Y_{(-i)} \tag{15}$$

kde $X_{(-i)}$ a $Y_{(-i)}$ znázorňují matici a vektor bez i-tého pozorování, tj.

$$X_{(-i)} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i-11} & x_{i-12} & \cdots & x_{i-1p} \\ x_{i+11} & x_{i+12} & \cdots & x_{i+1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{16}$$

$$Y_{(-i)} = [y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n]^T \tag{17}$$

Studentizovaná jackknife rezidua mají nulovou střední hodnotu, jednotkový rozptyl a jsou navzájem závislá a navíc mají Studentovo t-rozdělení se stupni volnosti $\nu = n - p - 1$. A právě na základě znalosti rozdělení těchto reziduí lze testovat odlehlá pozorování [26]. Za odlehlá lze posuzovat taková rezidua, jejichž absolutní hodnota je větší než kvantil t-rozdělení $t_{1-\alpha/2n}$.

Pro detekci odlehlých hodnot však budou využity spíše míry, k tomu určené, které využijí výše definovaných typů reziduí.

2.4.2 Míry detekce vlivných bodů

Již výše bylo zmíněno, že při detekci odlehlých hodnot v regresi, jsou tyto body označovány spíše jako vlivné body, které zároveň můžeme rozdělit na vlivné body ve směru osy x (*leverage points)* a vlivné body ve směru osy y (*outliers)*. Stejně tak existují míry detekující právě leverage points a outliers.

2.4.2.1 Detekce leverage points pomocí projekční matice H

Pro detekci leverage points lze využít projekční matici, kterou lze získat z dat jako $H = X(X^T X)^{-1} X^T$. Nenachází–li se v datech žádné odlehlé hodnoty ve směru osy x, tedy leverage points, pak by všechny diagonální prvky matice (h_{ii}) měly být přibližně stejné [27]. Je tedy nutné hledat vybočující hodnoty na diagonále projekční matice.

Pro projekční matici H také platí, že její stopa odpovídá počtu odhadovaných parametrů. Platí tedy

$$tr\ H = \sum_{i=1}^{n} h_{ii} = p \tag{18}$$

Pak by mělo platit, pokud se v datech nenachází leverage points

$$h_{ii} \approx \frac{p}{n} \tag{19}$$

Jako body podezřelé z odlehlosti ve směru osy x, lze považovat pozorování, pro která platí [20, 22]

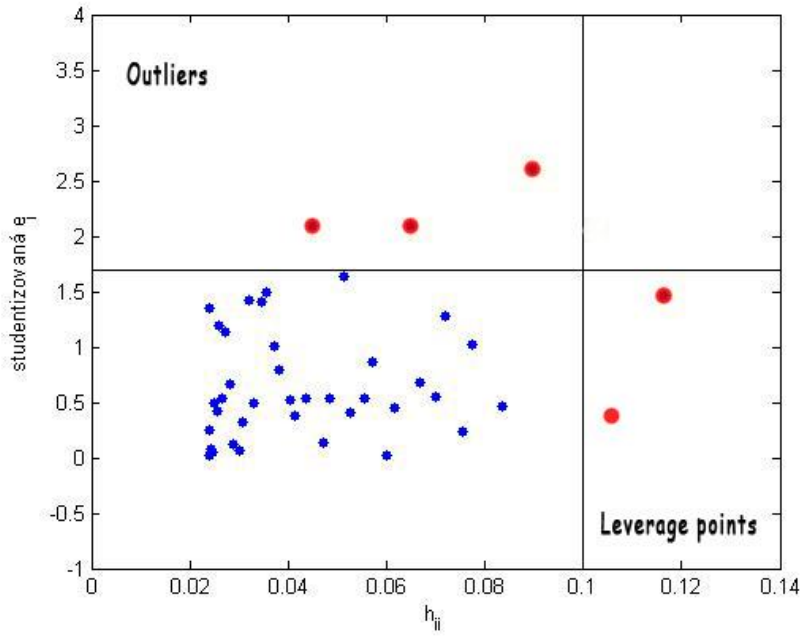
$$h_{ii} > \frac{2p}{n} \tag{20}$$

2.4.2.2 Williamsův graf

Pro detekci leverage points se využívá také nepřeberné množství grafických metod. V této práci bude zmíněna alespoň jedna tato metoda, a to tzv. *Williamsův graf,* kde je detekce leverage points opět založena na diagonálních hodnotách projekční matice H [22,23].

V tomto grafu jsou na ose x znázorněny diagonální prvky projekční matice a na ose y pak Studentizovaná jackknife rezidua v absolutní hodnotě. V grafu jsou pak uvedeny také mezní linie pro detekci jak leverage points, tak outliers. A to mezní linie pro outliers, tedy ve směru osy y: $y = t_{1-\alpha}(n-p)$, což značí $(1-\alpha)\%$ kvantil Studentova rozdělení s $n-p$ stupni volnosti a poté mezní linie pro leverage points, tedy ve směru osy x: $x = \frac{2p}{n}$.

Výsledný graf může vypadat následovně



Obrázek 2.4.2: Williamsův graf detekce leverage points

V *Obrázku 2.4.2* lze pak identifikovat leverage points vpravo od mezní hodnoty, tedy svislé přímky a následně též outliers nad mezní hodnotou, tedy vodorovnou přímkou.

2.4.2.3 Cookova vzdálenost

Cookova vzdálenost je často využívanou metodou pro identifikaci outliers v regresi. Tato metoda měří vliv i-tého pozorování na hodnotu odhadu vektoru β regresního modelu [27].

Cookova vzdálenost je definována

$$D_i = \frac{\left(\hat{Y} - \hat{Y}_{(-i)}\right)^T \left(\hat{Y} - \hat{Y}_{(-i)}\right)}{ps^2} \tag{21}$$

Tato vzdálenost udává změnu reziduálního součtu čtverců způsobenou vypuštěním i-tého pozorování.

Jinak lze Cookovu vzdálenost definovat také díky znalosti Studentizovaných jackknife reziduí a projekční matice H [22]

$$D_i = \frac{e_i^{(T)^2}}{p} \cdot \frac{h_{ii}}{1 - h_{ii}} \tag{22}$$

Orientačně platí, že je-li Cookova vzdálenost $D_i > 1$, lze detekovat i-té pozorování jako vlivný bod. Hodnotu D_i je však možné porovnávat také s kvantilem Fisherova rozdělení, a to konkrétně s kvantilem $F_\alpha(p, n-p)$. Outliers jsou tedy identifikovány, jestliže platí $D_i > F_\alpha(p, n-p)$.

2.4.2.4 Welsch-Kuhova vzdálenost

Tato metoda měří vliv i-tého pozorování nejen na hodnotu odhadu vektoru β, ale simultánně také na odhad parametru σ^2 [20].

Welsch-Kuhova vzdálenost je tak definována

$$WK_i = \frac{\left|\hat{y}_i - \hat{y}_{i(-i)}\right|}{s_{(-i)}\sqrt{h_{ii}}} \tag{23}$$

Stejně tak je možné tuto vzdálenost definovat pomocí Studentizovaných reziduí

$$WK_i = \left|e_i^{(T)}\right| \sqrt{\frac{h_{ii}}{1-h_{ii}}} \tag{24}$$

Vysoké hodnoty tohoto kritéria pak opět svědčí o výrazném vlivu i-tého pozorování. Tyto hodnoty je dle [25] vhodné srovnávat s mezní hodnotou

$$2\sqrt{\frac{p}{n}} \qquad \text{případně vhodněji s} \qquad 2\sqrt{\frac{p}{n-p}} \tag{25}$$

Tuto metodu lze nalézt též pod označením *DFFITS*, například ve zdroji [20] či jiných odborných materiálech.

3 Numerické experimenty

Výše uvedené metody identifikace odlehlých pozorování byly teoreticky popsány, ale pro lepší pochopení je vhodnější využít jednotlivé metody na konkrétních datech. K tomuto účelu tak budou provedeny numerické experimenty na generovaných datech v programu Matlab a následně také identifikovány odlehlé hodnoty. Provedení numerických experimentů by také mělo pomoci odhalit výhody a nevýhody jednotlivých metod, které budou použity.

3.1 Numerické experimenty jednorozměrných dat

V této části budou provedeny numerické experimenty pro jednorozměrná data. Z výše uvedených metod tak připadají v úvahu parametrické metody: Grubbsův test a Dean-Dixonův test. Jako doplňková metoda budou využity Boxploty, které graficky znázorní rozložení dat souboru.

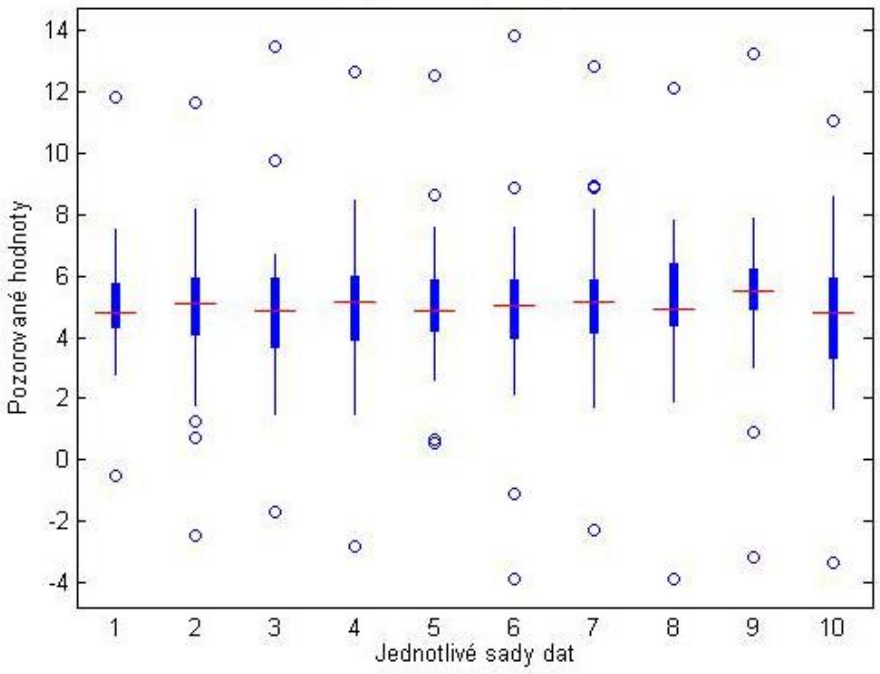
Obě parametrické metody jsou založené na předpokladu toho, že data pochází z normálního rozdělení, a tak je nutné při generování dat tuto podmínku dodržet.

3.1.1 Grubbsův test

Jak již bylo řečeno, tento test je vhodný pro větší soubory dat. Pro účely numerických experimentů bylo zvoleno několik sad dat, každá o 50 hodnotách. Samozřejmostí je variabilní počet hodnot, počet sad, ale také parametry normálního rozdělení, tedy střední hodnota a rozptyl.

Jako první konkrétní parametry generovaných dat byly zvoleny sady o 50 hodnotách. Těchto sad pak 10 při jednom kroku generování. Pro všechny tyto soubory dat byly zvoleny parametry normálního rozdělení $\mu = 5, \sigma^2 = 2$. Takto generovaná data pak byla záměrně „poškozena“ několika hodnotami, které mohou znázorňovat hrubé chyby při sběru dat nebo jejich následném zpracování. Toto poškození bylo provedeno pomocí libovolných hodnot, které byly přičteny nebo naopak odečteny od hodnot generovaných z normálního rozdělení. Zvoleny byly 4 různé hodnoty pro přičtení, a to 2 kladné a 2 záporné. Konkrétně byla zvolena poškození $8, 3, -3, -7$. V každé sadě se tak nachází 4 záměrně poškozené hodnoty, které by měly být identifikovány na základě Grubbsova testu pro odlehlé hodnoty.

Graficky lze tato generovaná data zobrazit pomocí Boxplotů následovně:



Obrázek 3.1.1: Boxploty numerického experimentu pro Grubbsův test

Na *Obrázku 3.1.1* jsou znázorněny mediány jednotlivých souborů dat červenou linií a pozorování podezřelá z odlehlosti pak modrými kolečky. Z charakteru generovaných dat tak v každém souboru vznikl jiný počet případně odlehlých hodnot. Dále bude tato grafická metoda porovnána s výsledky Grubbsova testu. Na data byla tato metoda aplikována opakovaně, dokud nebyly obě hodnoty testovacích kritérií T_1 a T_n vypočtené dle *vzorců (1)* menší než kritická hodnota T_α.

	Sady generovaných dat									
	1	2	3	4	5	6	7	8	9	10
Outliers	-0,531	-2,453	13,495	-2,793	12,553	-3,877	-2,280	-3,899	-3,158	-3,329
	11,800	11,659	-1,661	12,641		-1,092	12,814	12,139	13,204	11,075
			9,768			13,849			0,925	

Tabulka 3.1.1: Zjištěné outliery generovaných dat Grubbsovým testem

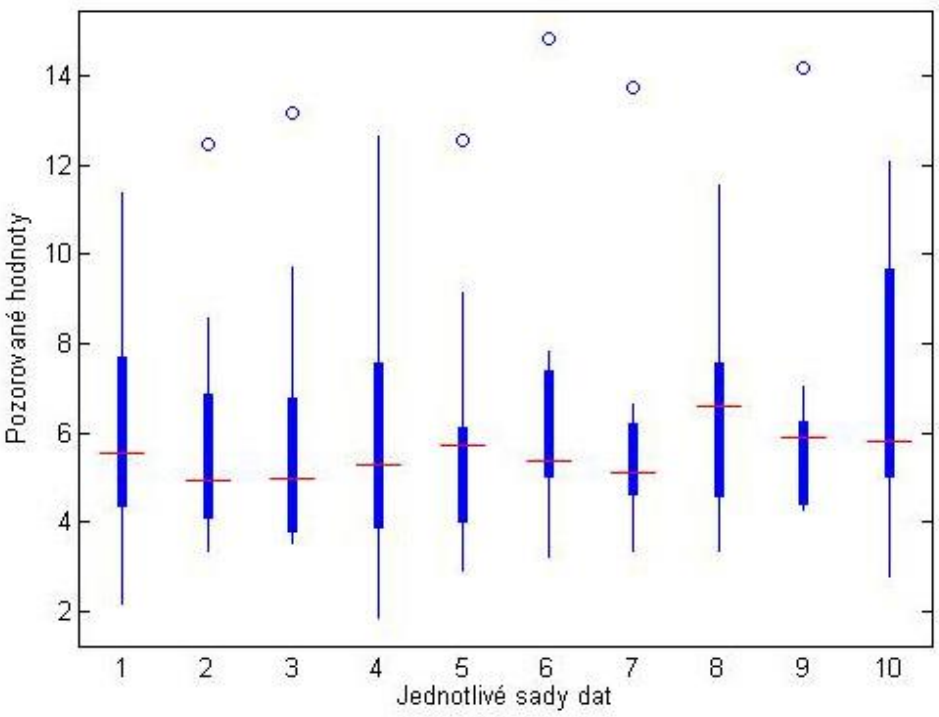
Z *Tabulky 3.1.1* je vidět, že ze 4 záměrně poškozených hodnot test odhalil v nadpoloviční většině pouze 2. Také lze říci, že v porovnání s výše zobrazeným grafem, je právě Grubbsův test poněkud flexibilnější, neboť obecně detekuje méně odlehlých hodnot v jednotlivých sadách, než Boxploty. Zatímco například v 5. sadě jsou z grafu znatelné 4 odlehlé hodnoty, Grubbsův test odhalil pouze jednu. Ve většině případů nebyly odhaleny jako odlehlé hodnoty ty, které byly poškozeny nižší hodnotou z dvojice zvolených pro přičtení (případně odečtení) ke generovaným datům.

3.1.2 Dean-Dixonův test

Tento test je opět závislý na předpokladu normálního rozdělení, ale na rozdíl od Grubbsova testu je tato metoda určena spíše pro menší rozsahy dat.

Opět bude tvořeno 10 souborů dat vždy s 10 pozorováními, což je maximální možný počet pro Dean-Dixonův test. S takto malým počtem pozorování není možné příliš poškozovat hodnoty, a tak budou poškozena pouze 2 pozorování, a to opět jedno více a jedno v menší míře. Konkrétně pak přičtením čísel 8 a 3. Model normálního rozdělení byl ponechán s parametry $\mu = 5, \sigma^2 = 2$.

S těmito parametry byly získány následující Boxploty



Obrázek 3.1.2: Boxploty numerického experimentu pro Dean-Dixonův test

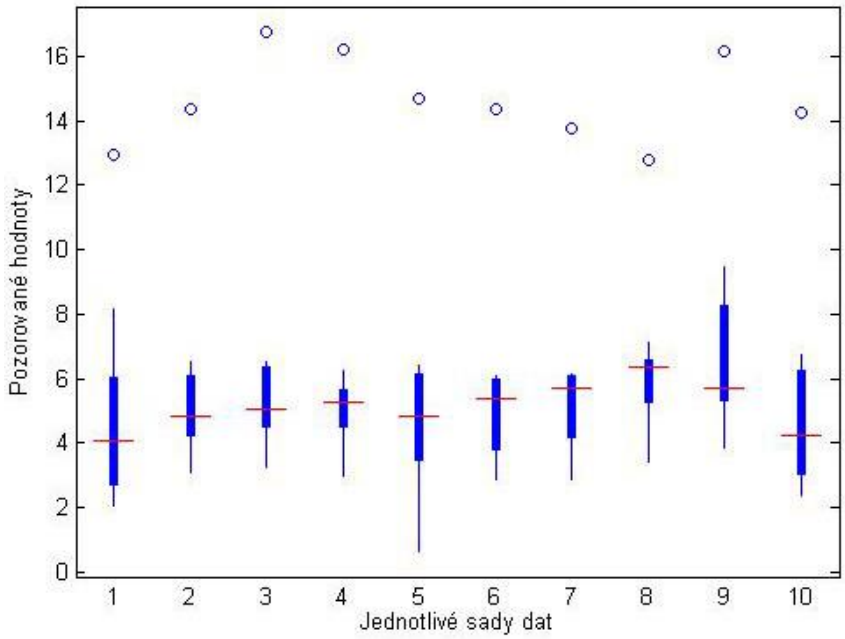
A také hodnoty Dean-Dixonova testu, který odhalil následující odlehlá pozorování.

	Sady generovaných dat									
	1	2	3	4	5	6	7	8	9	10
Outliers	-	12,447	-	12,645	-	14,812	13,750	-	14,178	-

Tabulka 3.1.2: Zjištěné outliery generovaných dat Dean-Dixonovým testem

Z těchto hodnot je patrné, že ačkoliv byly v souborech dat záměrně poškozeny 2 hodnoty, v některých sadách dat nebyly vůbec detekovány. V případě, že byla nějaká hodnota odhalena, vždy byla touto hodnotou ta s vyšším poškozením. Bohužel se ale také může stát, že poškozením souboru dat o 10 pozorováních dvěma hrubými chybami, došlo k porušení předpokladu normálního rozdělení.

Z čísel získaných při experimentech Dean-Dixonova testu bylo zjištěno, že metoda funguje spolehlivě, obsahují-li data jednu velmi výrazně odlehlou hodnotu. Toto tvrzení je možné doložit příkladem, kdy bylo poškozeno vždy jen jedno pozorování přičtením hodnoty 10.



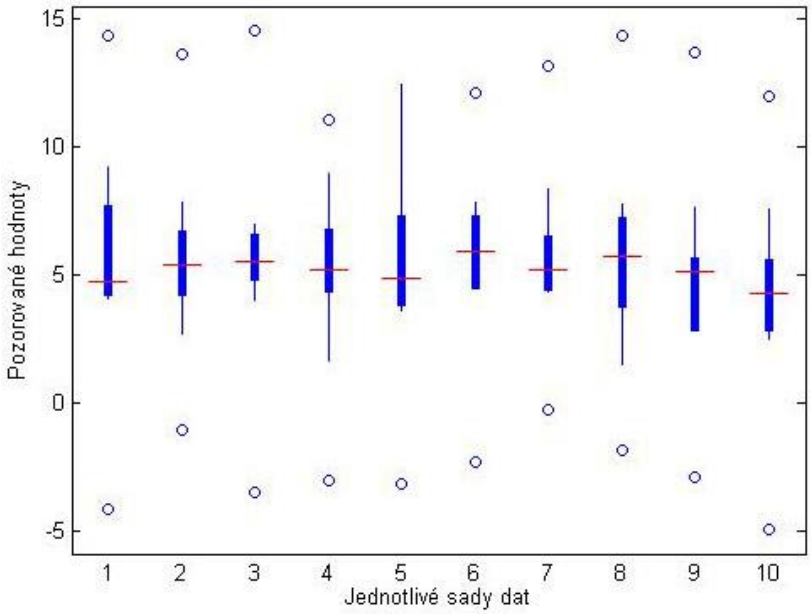
Obrázek 3.1.3: Boxploty pro Dean-Dixonův test s jedním poškozením souboru dat

	Sady generovaných dat									
	1	2	3	4	5	6	7	8	9	10
Outliers	12,933	14,350	16,750	16,222	14,692	14,377	13,752	12,800	16,160	14,254

Tabulka 3.1.3: Detekované odlehlé hodnoty Dean-Dixonovým testem

Problémem této metody může být fakt, že je založená na podílu vzdálenosti dvou sousedních hodnot a rozpětí celého souboru dat. Bude-li totiž soubor záměrně poškozen jedním pozorováním odlehlým ve směru kladných hodnot a druhým ve směru záporných hodnot, navyšuje se zmíněné rozpětí souboru, což zapříčiňuje snižování testové statistiky a následné případné zamítnutí hypotézy o odlehlosti hodnot.

Tato vlastnost může být demonstrována v následujícím příkladu. Bude ponechán model se stejnými parametry, tedy 10 souborů hodnot vždy po 10 pozorováních s předpokladem normálního rozdělení s parametry $\mu = 5, \sigma^2 = 2$. Poškozeny nyní budou 2 hodnoty, a to přičtením čísla 8 a také jeho odečtením.



Obrázek 3.1.4: Boxploty pro Dean-Dixonův test se dvěma poškozeními

Zatímco Boxploty odhalily odlehlé hodnoty v celkem devíti sadách dat, Dean-Dixonův test pouze v polovině případů. V pěti sadách dat tak vyhodnotil hodnoty jako odpovídající, a tedy žádné odlehlé.

	Sady generovaných dat									
	1	2	3	4	5	6	7	8	9	10
Outliers	-4,118	-	-3,507	-	-3,144	-2,264	-	-	-	-4,906
	14,345	-	14,542	-	12,463	12,093	-	-	-	11,999

Tabulka 3.1.4: Detekované odlehlé hodnoty Dean-Dixonovým testem

Výše uvedenými numerickými experimenty jednorozměrných metod, byly zjištěny hlavní výhody i nevýhody Grubbsova a Dean-Dixonova testu. Tyto metody detekování outliers jsou jednodušší na implementaci algoritmu pro zvolená data. Také lze říci, že nejsou náročné na statistický aparát potřebný k aplikaci metody, neboť jsou založeny na testovací statistice, kterou je nutné vždy porovnat s kritickou hodnotou, aby bylo možné vždy přijmout či zamítnout hypotézu odlehlého pozorování.

Jako hlavní nevýhodu lze chápat předpoklad normálního rozdělení. Zvláště pro Dean-Dixonův test, kdy soubor dat obsahuje nejvýše 10 pozorování, je těžké s jistotou tvrdit, že zkoumaná data odpovídají normálnímu rozdělení, zvláště pak, pokud obsahují odlehlou hodnotu.

Pokud tedy data neodpovídají normálnímu rozdělení, je možné aplikovat Boxplot, jako neparametrickou metodu identifikace outliers. Zde je ovšem zase velmi důležité dávat pozor na výše zmíněné efekty, které se mohou objevovat v souborech dat, tedy *masking effect* a *swamping effect* v jednorozměrném provedení.

3.2 Numerické experimenty vícerozměrných dat

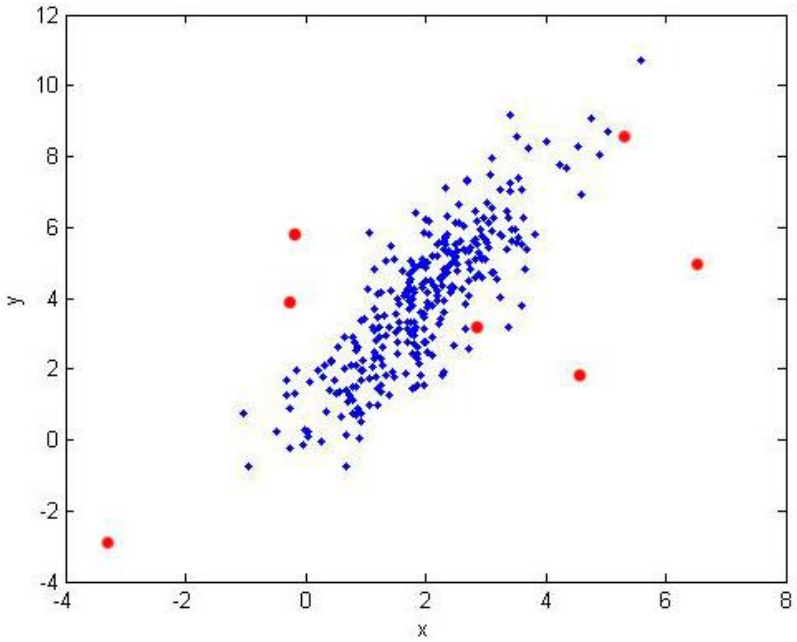
3.2.1 Mahalanobisova vzdálenost

Numerické experimenty Mahalanobisovy vzdálenosti budou prováděny na datech pocházejících z dvourozměrného normálního rozdělení. Tato data jsou náhodně generována pomocí výpočetního prostředí Matlab, kde byl využit příkaz *mvnrnd*. Tento příkaz umožňuje generovat data dvourozměrného normálního rozdělení s předem definovanými středními hodnotami a kovarianční maticí.

Pro účely numerických experimentů byly zvoleny parametry dvourozměrného normálního rozdělení:

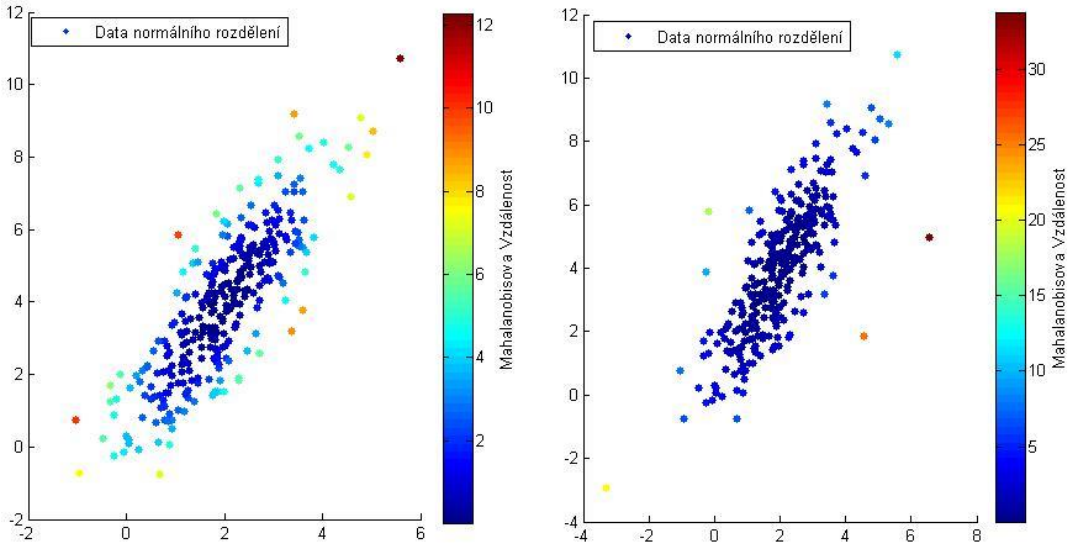
$$\boldsymbol{\mu} = (2,4)^T \quad , \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1{,}6 \\ 1{,}6 & 4 \end{pmatrix}$$

S těmito parametry bylo generováno 300 hodnot a následně sedm z nich záměrně poškozeno, jako demonstrace hrubých chyb v datech. Porušeny byly dvě hodnoty pouze ve směru osy x, dvě hodnoty pouze ve směru osy y a tři hodnoty v obou směrech zároveň. Budou-li tato data zobrazena v grafu, budou lépe vidět poškozené hodnoty.



Obrázek 3.2.1: Generovaná data spolu s poškozenými hodnotami

Na *Obrázku 3.2.1* jsou červeně znázorněna záměrně poškozená data. Ovšem hned dva tyto body se vizuálně nachází velmi blízko většině dat a je tak možné, že nebudou hodnoceny jako odlehlé. Při posuzování odlehlosti dat pomocí Mahalanobisovy vzdálenosti byly tyto vzdálenosti vypočteny pro všechny body. Pozorování podezřelá z odlehlosti jsou pro hodnotitele ta, která nabývají nejvyšších čísel Mahalanobisovy vzdálenosti. Tyto hodnoty lze v Matlabu spočítat pomocí příkazu *mahal (X,Y)*. Pro lepší názornost pak mohou být data vykreslena v barevné škále odstupňované podle narůstající Mahalanobisovy vzdálenosti.



Obrázek 3.2.2: Mahalanobisova vzdálenost generovaných dat (vlevo) a poškozených dat (vpravo)

V *Obrázku 3.2.2* je vidět, že nejvzdálenější hodnotou z generovaných dat (zobrazené na grafu vlevo) dle Mahalanobisovy vzdálenosti je bod zabarvený rudou barvou v pravém horním rohu grafu. Tento bod dosahuje Mahalanobisovy vzdálenosti 12,25 a vzhledem k parametrům vypočteným ze všech dat mohou být jako outliers hodnoceny i ostatní červeně zabarvené body, které jsou ve vzdálenosti 10,03 či 9,84.

Zatímco v případě, kdy jsou některá pozorování záměrně poškozena přičtením nebo odečtením hodnot 3 až 4, dochází k větším odstupům hodnot od zbytku dat. Tímto konkrétním poškozením souboru došlo k nárůstu maximální hodnoty Mahalanobisovy vzdálenosti na 33,74, což odpovídá rudému bodu situovanému vpravo uprostřed (graf vpravo). Zde je vidět, že hodnoty původně vygenerované se i s Mahalanobisovou vzdáleností kolem čísla 12 nemusí hodnotit jako odlehlé. Pomocí této metody tak byly detekovány všechny záměrně poškozené hodnoty kromě jedné. Budou-li totiž zobrazeny vzdálenosti bez přičtených poškozujících hodnot a poté s nimi, lze odhalit, že mezi největšími Mahalanobisovými vzdálenosti jsou právě ty, které byly záměrně poškozeny. Odhaleny však nejsou všechny.

				Vzdálenosti seřazené od nejvyšší	
		MV - generovaná data	MV - poškozená data	MV - generovaná data	MV - poškozená data
Poškozená data	1	0,283	**33,739**	12,247	**33,739**
	2	4,063	**18,372**	10,025	**25,688**
	3	4,610	**9,727**	9,835	**20,925**
	4	1,232	3,077	8,961	**18,372**
	5	0,970	**8,213**	8,755	11,351
	6	2,026	**20,925**	8,629	**9,727**
	7	0,387	**25,688**	8,244	**8,213**
	8	0,129	0,122	7,827	7,983
	9	12,247	11,351	7,539	7,407
	10	7,214	6,625	7,214	7,042

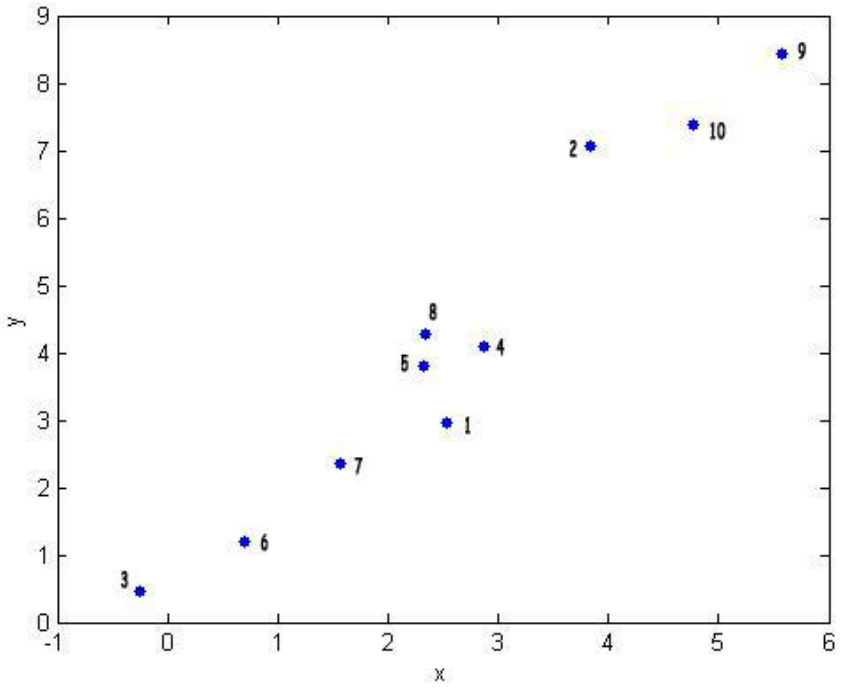
Tabulka 3.2.1: Mahalanobisovy vzdálenosti prvních deseti bodů

V *Tabulce 3.2.1* je zobrazeno prvních deset bodů a jejich Mahalanobisovy vzdálenosti, přičemž prvních sedm bodů bylo záměrně poškozeno. V pravé části tabulky jsou pak všechny hodnoty souboru seřazeny od největší vzdálenosti. Nyní už je pouze na pozorovateli, které hodnoty na základě Mahalanobisovy vzdálenosti vyhodnotí jako odlehlé. Budou-li jako odlehlé brány hodnoty převyšující číslo 18, obsahuje soubor s upravenými pozorováními 4 odlehlé hodnoty. Všechny čtyři hodnoty pak byly záměrně poškozeny. Z tabulky lze vyčíst, že těmito hodnotami byly první, druhá, šestá a sedmá.

Hlavní nevýhodou této metody je absence jakéhokoliv rozhodovacího kritéria. Je tak výhradně na pozorovateli, zda hodnoty podezřelé z odlehlosti vyhodnotí jako outliers. Tato metoda je tak spíše doplňující grafická a bylo by vhodné, doplnit ji před rozhodnutím ještě nějakou jinou, která by potvrdila nebo naopak vyvrátila domněnky pozorovatele o odlehlosti.

3.2.2 Metoda KDIST a MeanDIST

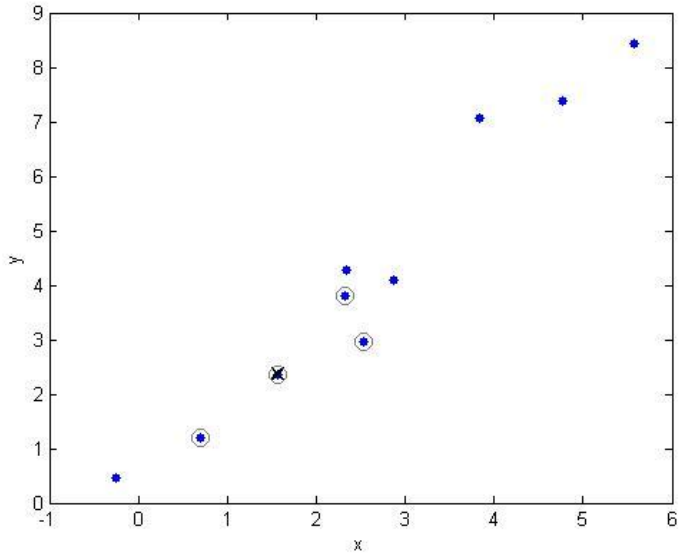
Tyto metody využívají princip *k*-nejbližších sousedů, přičemž hlavním rozdílem mezi nimi je hodnota vzdálenosti, která je dále využívána. V případě metody *KDIST*, je dále každému bodu přiřazena hodnota maximální vzdálenosti z *k*-nejbližších sousedů od tohoto bodu. V případě *MeanDIST,* je přiřazenou hodnotou průměr vzdáleností všech *k*-nejbližších sousedů zvoleného bodu. Pro tyto dvě metody bylo vygenerováno 10 pozorování, na kterých bude demonstrován princip metod. Tato pozorování jsou znázorněna na následujícím *Obrázku 3.2.3*. Spolu s nimi jsou na obrázku uvedena pořadí, ve kterém byly hodnoty generované, pro lepší orientaci v následných výpočtech.



Obrázek 3.2.3: Data generovaná pro metody KDIST a MeanDIST

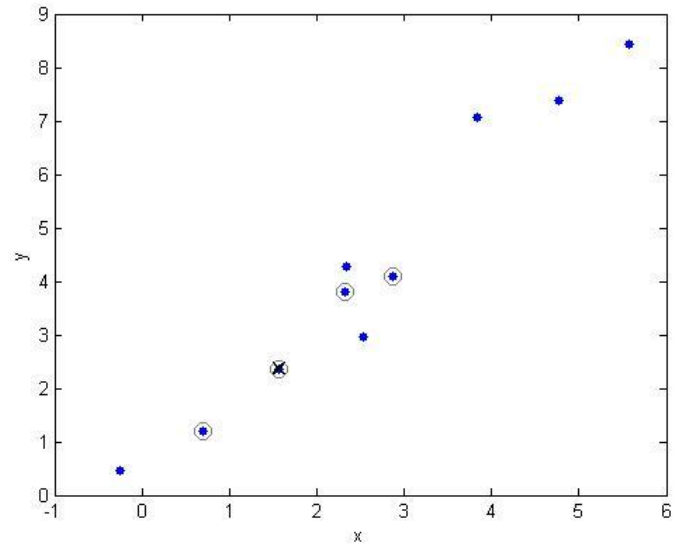
Nyní je nutné pro každý bod nalézt *k*-nejbližších sousedů. V tomto případě bylo zvoleno $k = 3$. Je několik možností, jak měřit vzdálenost mezi body. V této práci byla zvolena Euklidovská vzdálenost a následně Mahalanobisova. Pomocí obou typů vzdáleností byly pro každý bod určeny vzdálenosti od *3*-nejbližších sousedů a následně vypočteny hodnoty *KDIST* a *MeanDIST*.

V matlabovském programu jsou vyobrazeny sousedé pro každý bod. Zde bude pro názornost uveden alespoň jeden takový, kde zvolený bod je označen černým křížkem a jeho *3*-nejbližší sousedé pak šedým kroužkem.



Obrázek 3.2.4: 3-nejbližší sousedé sedmého bodu (Euklidovská vzdálenost)

V *Obrázku 3.2.4* jsou znázorněny 3 nejbližší sousedé sedmého bodu v pořadí pomocí Euklidovské vzdálenosti. Využije-li se však Mahalanobisova vzdálenost, dochází v několika případech k nalezení odlišných sousedních bodů. Například budou-li hledaní sousedé opět sedmého bodu v pořadí, ale pomocí Mahalanobisovy vzdálenosti, při využití kovarianční matice, pomocí které byla data generována, metoda nalezne jiné sousední body znázorněné v *Obrázku 3.2.5.*



Obrázek 3.2.5: 3-nejbližší sousedé sedmého bodu (Mahalanobisova vzdálenost)

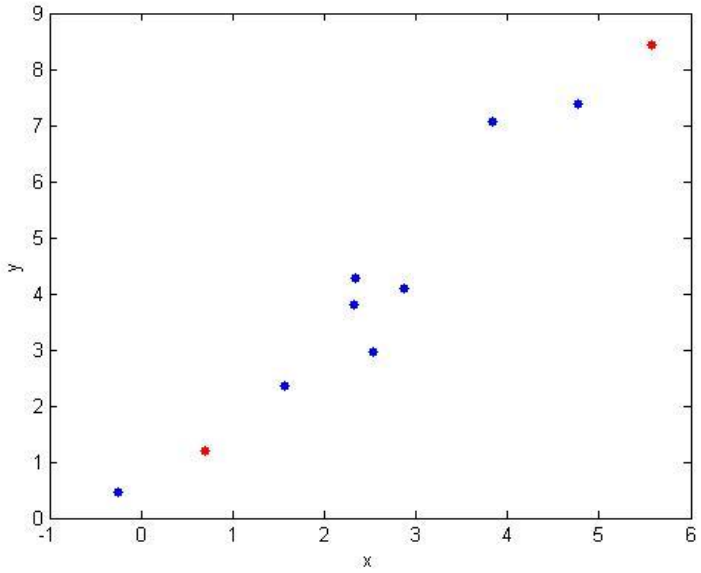
Odlišné budou také získané hodnoty *KDIST* a *MeanDIST*. Ty jsou pro srovnání zobrazeny v následující *Tabulce 3.2.2*.

	Euklidovská vzdálenost		Mahalanobisova vzdálenost	
Pořadí bodu	KDIST	MeanDIST	KDIST	MeanDIST
1	1,184	1,069	2,076	1,656
2	3,142	2,114	2,385	2,134
3	3,738	2,521	2,578	1,974
4	1,184	0,782	1,327	1,191
5	0,892	0,655	1,127	0,950
6	2,551	1,738	1,868	1,359
7	1,629	1,410	1,297	1,101
8	1,342	0,786	1,791	1,401
9	5,129	2,889	3,089	2,094
10	3,815	2,042	2,414	1,697

Tabulka 3.2.2: Vypočtené hodnoty KDIST a MeanDIST (Euklidovská i Mahalanobisova vzdálenost)

Na základě těchto vzdáleností byly následně vypočteny diference mezi seřazenými hodnotami, které jsou důležité pro výpočet mezní hodnoty, od které se body v souboru dat hodnotí jako možné outliers. Při výpočtu této mezní hodnoty je velmi důležitá volba parametru t, neboť udává, jak se změní mezní hranice v porovnání s maximální deferencí hodnot *KDIST* či *MeanDIST.*

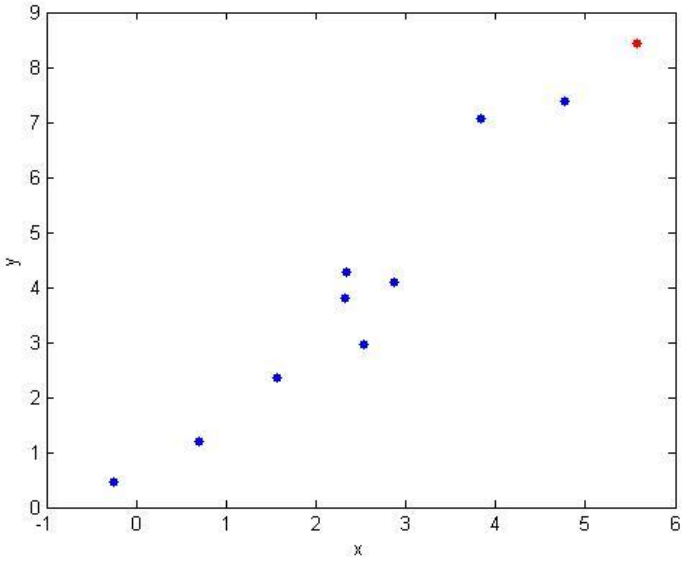
Metodou *KDIST* založenou na hledání sousedů Euklidovskou vzdáleností byly při volbě parametru $t = 0{,}5$ až $0{,}7$ nalezeny právě 2 body podezřelé z odlehlosti, a to konkrétně body šest a devět zobrazené na následujícím *Obrázku 3.2.6*.



Obrázek 3.2.6: Detekované outliers metodou KDIST (Euklidovská vzdálenost)

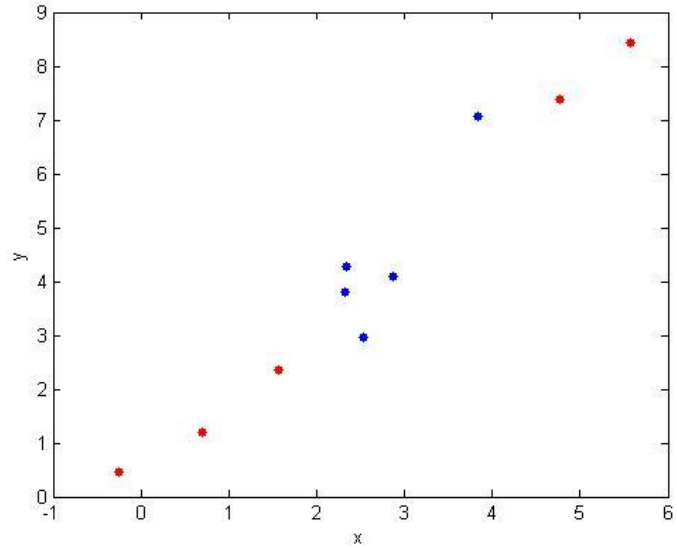
Právě na *Obrázku 3.2.6* je možné vidět, že může nastat problém s některým z efektů. Zároveň zde nastává situace, kdy vzdálenějším bodům byla přidělena podobná hodnota *KDIST* a bod číslo 3 (na *Obrázku 3.2.6* vlevo dole) nebyl vyhodnocen jako odlehlý, zatímco bližší bod číslo 6 ano.

Při volbě parametru $t = 0{,}75$ až $0{,}95$ je metodou nalezen pouze jeden outlier, a to bod číslo 9 zobrazený na následujícím *Obrázku 3.2.7.*



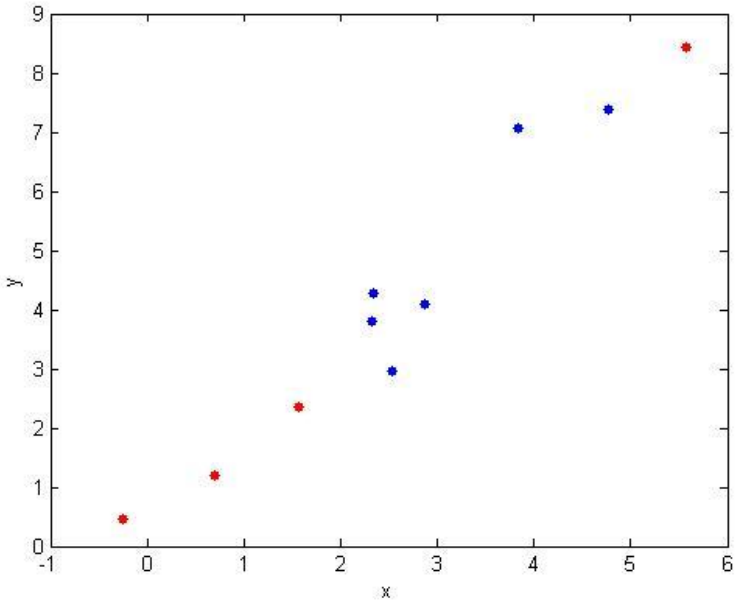
Obrázek 3.2.7: Detekované outliers metodou KDIST (Euklidovská vzdálenost)

Metodou *MeanDIST* založenou na hledání sousedů Euklidovskou vzdáleností bylo při volbě parametru $t = 0{,}7$ nalezeno až 5 bodů podezřelých z odlehlosti.



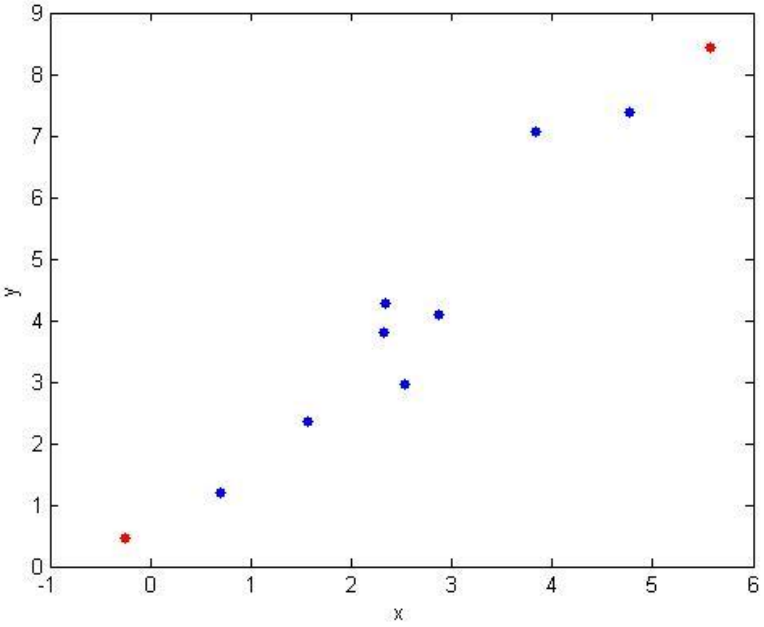
Obrázek 3.2.8: Detekované outliers metodou MeanDIST (Euklidovská vzdálenost)

Při volbě parametru $t = 0{,}75$ až $0{,}8$ jsou metodou *MeanDIST* nalezeny outliers čtyři, a to tak, že v pořadí desátý bod není detekován jako odlehlý.



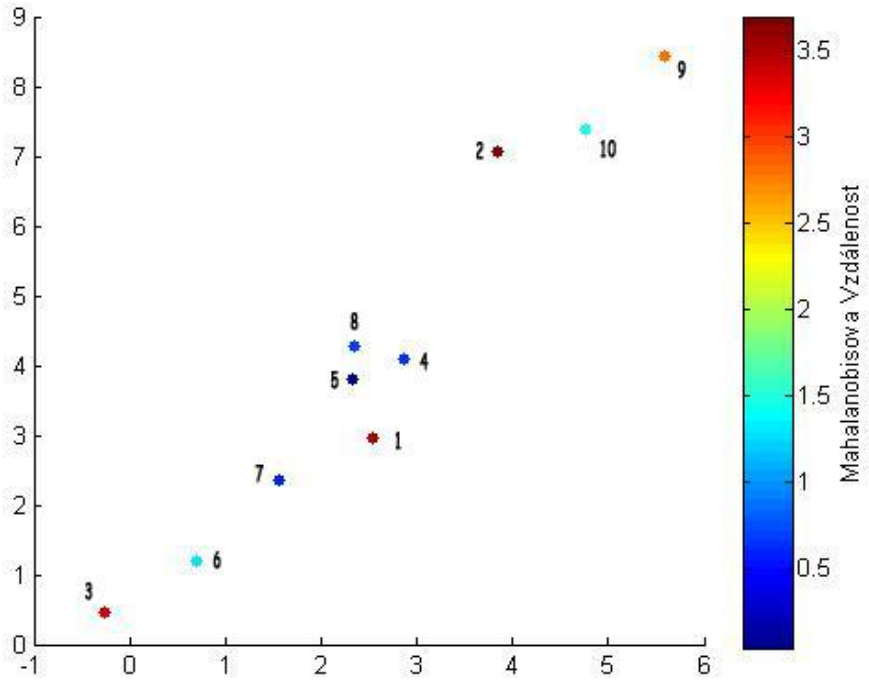
Obrázek 3.2.9: Detekované outliers metodou MeanDIST (Euklidovská vzdálenost)

Při volbě parametru $t = 0{,}85$ až $0{,}9$ jsou metodou nalezeny pouze 2 outliers. Zobrazeny jsou na následujícím *Obrázku 3.2.10.*



Obrázek 3.2.10: Detekované outliers metodou MeanDIST (Euklidovská vzdálenost)

Zatím byly prováděny numerické experimenty a detekce outliers metodami *KDIST* a *MeanDIST* za pomoci Euklidovské vzdálenosti, ale jak již bylo výše poznamenáno, je možné využít i jiné vzdálenosti, jako například Mahalanobisovu. Rozdíly v hodnotách *KDIST* a *MeanDIST* u těchto vzdáleností jsou zaznamenány již v *Tabulce 3.2.2.* Nyní tak bude tento rozdíl zhodnocen také z hlediska detekovaných odlehlých bodů. Jelikož je výpočet Mahalanobisovy vzdálenosti založen na předem definovaných parametrech mnohorozměrného rozdělení, bude pro tuto vzdálenost využita kovarianční matice, pomocí které byla data generována. Budou-li generovaná data barevně zobrazena dle Mahalanobisovy vzdálenosti, bude možné odhadovat odlehlé body.



Obrázek 3.2.11: Mahalanobisovy vzdálenosti generovaných dat pro metody KDIST a MeanDIST

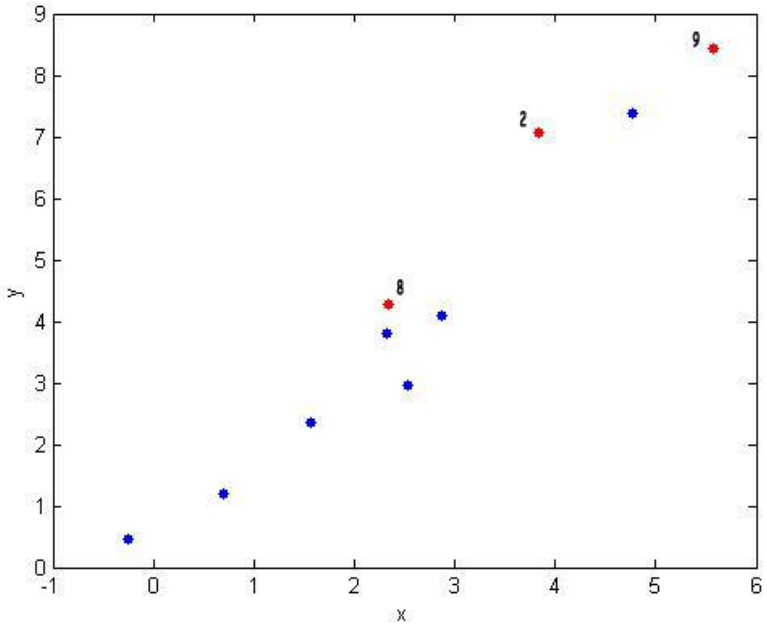
Díky *Obrázku 3.2.11* by mohly být jako odlehlé body hodnoceny body číslo 1,2,3 a možná také 9, což jsou pozorování hodnocená jako odlehlá odlišně od předchozích, které byly počítány pomocí Euklidovské vzdálenosti.

Budou-li nyní na tato data aplikovány metody detekce outliers *KDIST* a *MeanDIST* založené na Mahalanobisových vzdálenostech, budou získány následující hodnoty bodů podezřelých z odlehlosti. Body jsou zapsány v následující Tabulce podle svého pořadí generování.

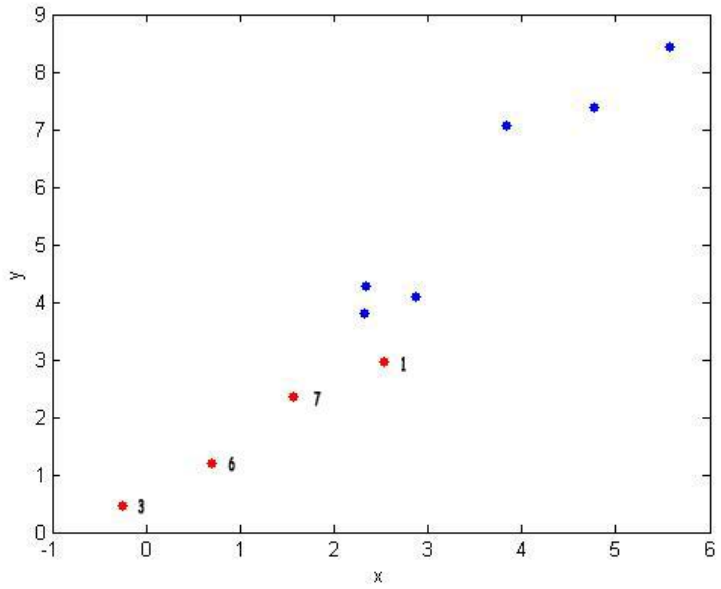
t	KDIST	MeanDIST
0,50	2,8,9	1,3,6,7
0,55	2,8,9	1,3,6
0,60	2,8,9	1,3
0,65	8,9	1,3
0,70	8,9	1,3
0,75	8,9	1,3
0,80	8,9	1,3
0,85	8,9	1,3
0,90	8,9	1,3
0,95	9	3

Tabulka 3.2.3: Detekované outliers metodami KDIST a MeanDIST (Mahalanobisova vzdálenost)

V této tabulce je vidět, že každá z metod detekovala jiné odlehlé body. Zatímco metodou *KDIST* jsou za odlehlé považované body v pravé horní části souboru dat, metoda *MeanDIST* odhalila outliers spíše v levé spodní části. Tato skutečnost je zobrazena na následujících *Obrázcích 3.2.12* a *3.2.13.*



Obrázek 3.2.12: Detekované outliers metodou KDIST, $t = 0,5$ až $0,6$ (Mahalanobisova vzdálenost)



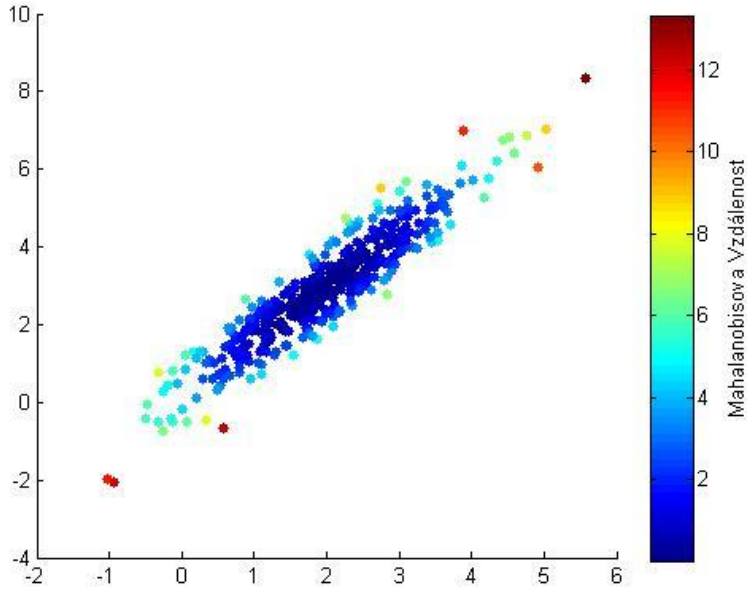
Obrázek 3.2.13: Detekované outliers metodou MeanDIST, $t = 0,5$ (Mahalanobisova vzdálenost)

V rámci těchto dvou metod je hned několik důležitých volených parametrů. Nejprve je velmi důležité, jaká vzdálenost bude zvolena pro hledání *k*-nejbližších sousedů. Rozdíl mezi detekovanými odlehlými body v závislosti na volbě vzdálenosti je patrný výše. Dalším důležitým parametrem je volba t, protože s narůstající hodnotou může identifikovat menší množství outliers. V neposlední řadě je také důležité, zda pro detekci bude využita metoda *KDIST*, založená na vzdálenosti *k*-nejbližšího souseda, nebo metoda *MeanDIST,* založená na průměru vzdáleností *k*-nejbližších sousedů zvoleného bodu.

Problémem těchto metod je také výskyt nějakého efektu dat, který byl popsán v teoretické části práce (*Masking effect, Swamping effect*). U generovaných dat pro tyto metody nelze s jistotou vyloučit přítomnost nějakého tohoto efektu, také z důvodu malého počtu generovaných dat. Proto budou metody aplikovány na větší soubor dat generovaný z dvourozměrného normálního rozdělení s parametry

$$\boldsymbol{\mu} = (2,3)^T \quad , \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1{,}3 \\ 1{,}3 & 2 \end{pmatrix}.$$

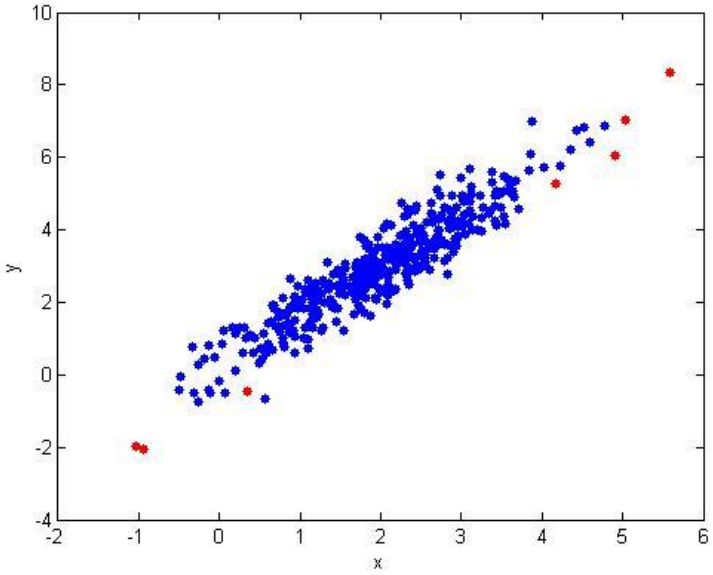
Generováno je celkem 400 hodnot, které je možné znázornit v následujícím *Obrázku 3.2.14*, udávající také Mahalanobisovy vzdálenosti všech bodů.



Obrázek 3.2.14: Mahalanobisovy vzdálenosti generovaných dat pro metody KDIST a MeanDIST

Z tohoto obrázku by mohlo být identifikováno přibližně 6 odlehlých hodnot. Nyní budou na tato data aplikovány metody *KDIST* a *MeanDIST*. Nejdříve založené na hledání *5*-nejbližších sousedů pomocí Mahalanobisovy vzdálenosti a následně pomocí Euklidovské vzdálenosti.

Zatímco metoda *KDIST* (Mahalanobisova vzdálenost) odhalila různý počet outliers v závislosti na volbě parametru t, metoda *MeanDIST* detekovala pro všechny hodnoty parametru t vždy jen jeden odlehlý bod.



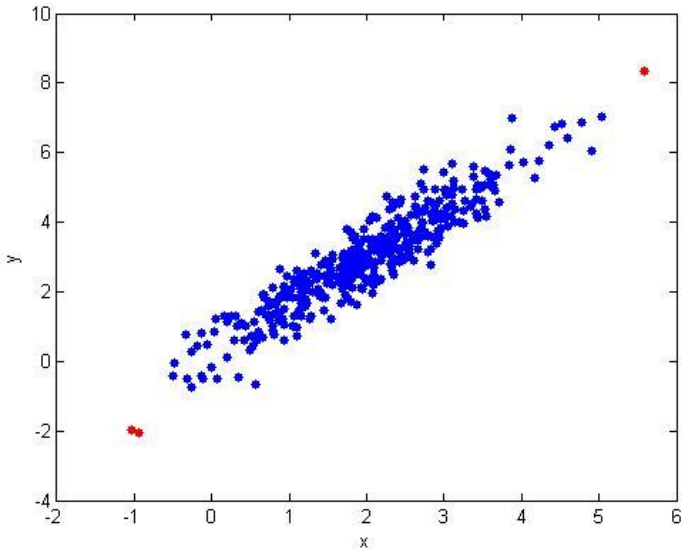
Obrázek 3.2.15: Detekované outliers metodou KDIST, $t = 0,5$ (Mahalanobisova vzdálenost)

Počet detekovaných outliers v závislosti na volbě parametru t metodou *KDIST* je zapsán v následující *Tabulce 3.2.4*.

t	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
Počet outliers	7	6	6	6	6	5	4	4	3	2

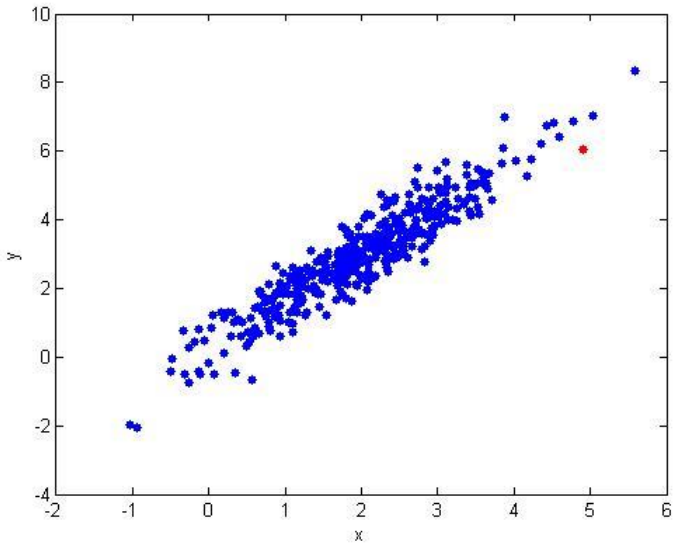
Tabulka 3.2.4: Počet detekovaných outliers metodou KDIST (Mahalanobisova vzdálenost)

Při volbě parametru $t = 0{,}9$ jsou tak detekované 3 odlehlé hodnoty zobrazené na následujícím *Obrázku 3.2.16*.



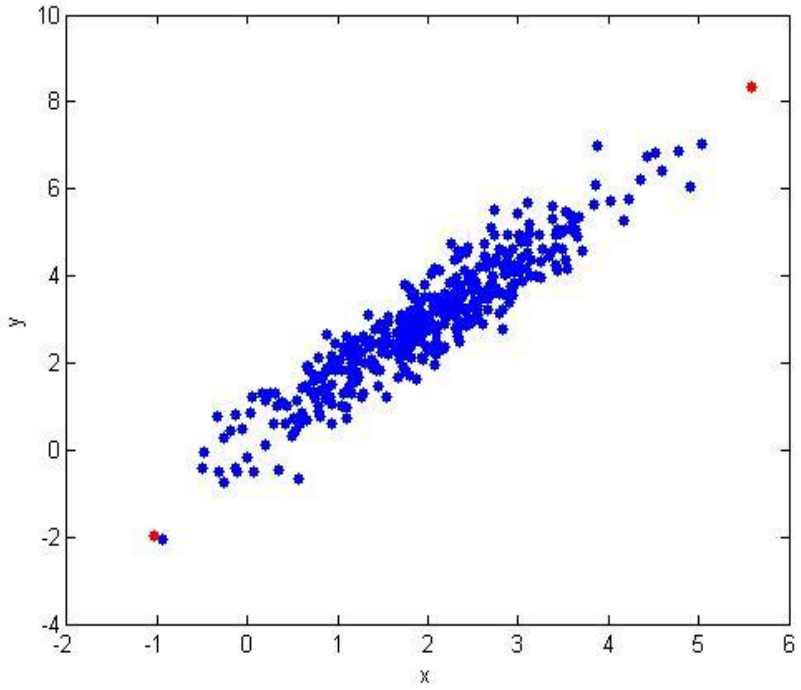
Obrázek 3.2.16: Detekované outliers metodou KDIST, $t = 0,9$ (Mahalanobisova vzdálenost)

Jak již bylo zmíněno výše, metodou *MeanDIST* byl jako odlehlý identifikován pouze jediný bod, a to bez ohledu na změnu voleného parametru t.



Obrázek 3.2.17: Detekovaný outlier metodou MeanDIST (Mahalanobisova vzdálenost)

Využitím Euklidovské vzdálenosti namísto Mahalanobisovy se změní především počet detekovaných outliers metodou *KDIST*, kde pro volený parametr $t = 0{,}5$ případně $0{,}55$ jsou identifikovány 2 odlehlé body zobrazené v následujícím *Obrázku 3.2.18*.

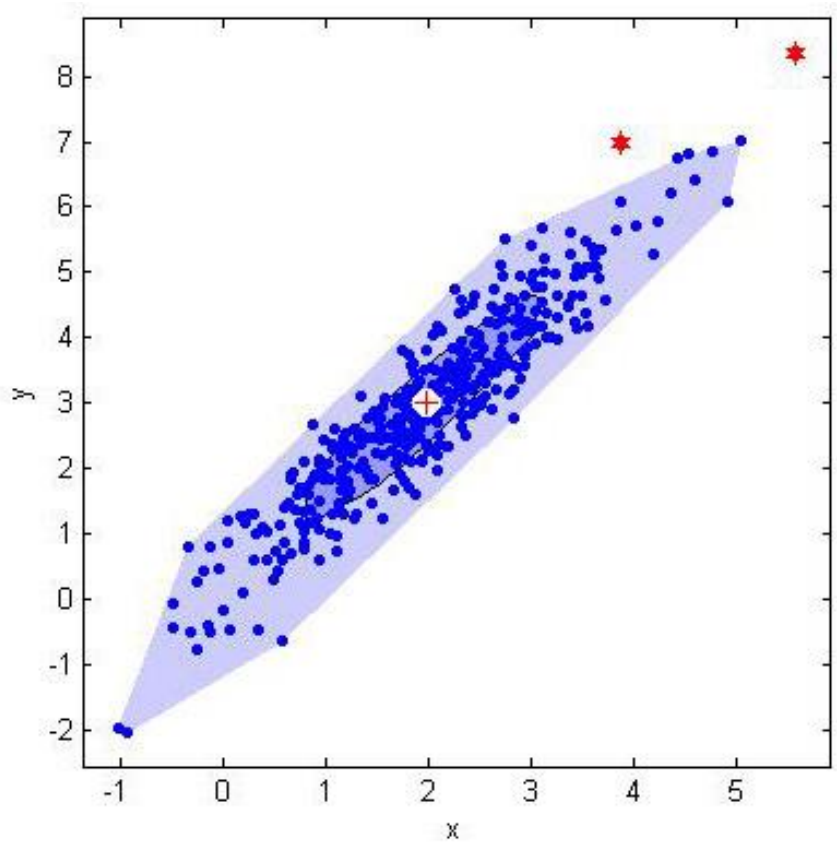


Obrázek 3.2.18: Detekované outliers metodou KDIST, $t = 0,5$ (Euklidovská vzdálenost)

Volbou parametru t vyšší než 0,55 je získán pouze jeden odlehlý bod, a to ten, který se na *Obrázku 3.2.18* nachází v levém dolní rohu.

Metodou *MeanDIST,* která hledá sousedy jednotlivých bodů pomocí Euklidovských vzdáleností, jsou získány dva odlehlé body bez ohledu na volbu parametru t. Takže pro každou volbu parametru t v rozmezí 0,50 až 0,95 jsou nalezeny odlehlé body totožné s metodou *KDIST*, $t = 0{,}5$, jejíž outliers jsou znázorněny v *Obrázku 3.2.18.*

Bude-li v závěru numerických experimentů metod *KDIST* a *MeanDIST* ještě využita grafická metoda bagplot, budou získány dva odlehlé body. Výsledný bagplot je zobrazený na následujícím *Obrázku 3.2.19.* Nejen díky tomuto bagplotu lze konstatovat, že bod v pravém horním rohu souboru dat může být odlehlým bodem, a tak by bylo vhodné důkladně zvážit, zda hodnotu ponechat k dalšímu analyzování nebo ji raději vyloučit ze souboru, aby neskreslovala případné výsledky statistických analýz.



Obrázek 3.2.19: Bagplot generovaných dat pro metody KDIST a MeanDIST

Velkou výhodou metod *KDIST* a *MeanDIST* je fakt, že jsou to metody neparametrické a je tedy možné využít jejich aparát na jakákoliv dvourozměrná data, kde není přítomen žádný nežádoucí efekt. Také je možné libovolně měnit počet nejbližších sousedů, pomocí kterých budou potřebné hodnoty počítány. Stejně tak je zcela na pozorovateli volba typu vzdálenosti mezi sousedními body.

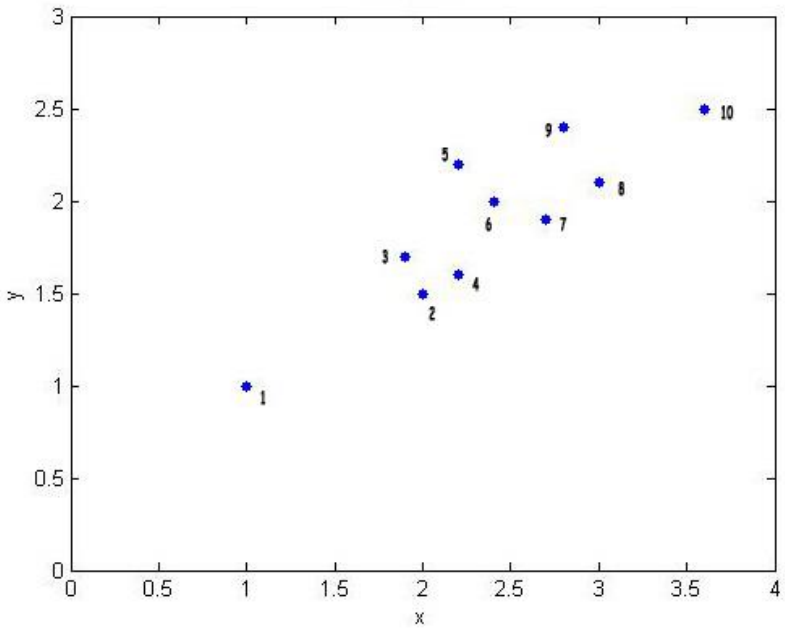
Avšak metody mají také nevýhody. Jedna z nich spočívá v již zmíněném problému, kdy ze dvou blízkých hodnot byla jako outlier detekována pouze jedna, zatímco druhá ne. Tento jev lze pozorovat například na *Obrázku 3.2.6* nebo *3.2.18.* Toto je způsobeno malou diferencí podobně vzdálených bodů, čímž jeden z nich není odhalen. Vhodné by proto bylo již detekované outliers vyřadit ze souboru dat a zvolenou metodu aplikovat opakovaně.

Další nevýhodou se stává volba parametru t, na kterém závisí počet detekovaných odlehlých hodnot. Je tak velmi těžké tento parametr zvolit vhodně.

3.2.3 Local outlier factor

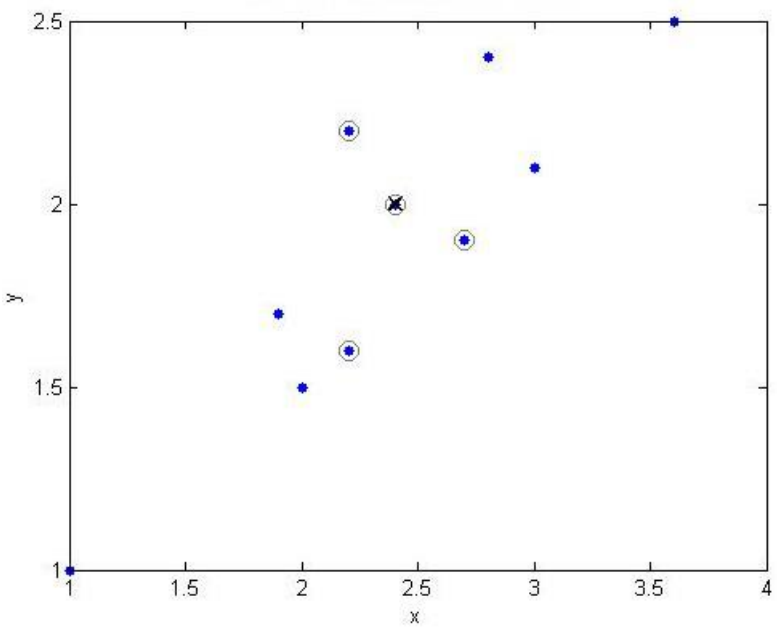
Tato metoda je založená na porovnávání hustoty sousedů v okolí jednotlivých bodů, přičemž odlehlými hodnotami jsou ty, jejichž hustota sousedů je nižší než ostatních dat. Hustota bodu je pak posuzována pomocí *k*-nejbližších sousedů.

Pro účely numerických experimentů byl zvolen soubor dat o velikosti deseti pozorování, na kterých bude demonstrován princip metody. Tato pozorování jsou zobrazena v *Obrázku 3.2.20* a zároveň jsou očíslována pro lepší orientaci v následujících výpočtech.



Obrázek 3.2.20: Data pro detekci outliers metodou LOF

Jelikož se jedná o metodu založenou na hustotě sousedů jednotlivých pozorování, lze očekávat, že případné odlehlé hodnoty se budou vyskytovat na okraji souboru dat. Aby však mohlo být rozhodnuto na základě metody LOF, je nutné nejdříve zjistit *k*-nejbližší sousedy jednotlivých bodů a jejich vzdálenosti. V práci byl zvolen parametr $k = 3$. Budou tak hledaní vždy 3 sousedé jednotlivých pozorování. Pro lepší názornost bude ukázán zvolený bod ze souboru pozorování a jeho 3 nejbližší sousedé. Tento bod je značen černým křížkem a jeho sousedé jsou pak v šedém kruhu. Vzdálenost mezi nimi byla počítána v programu Matlab jako Euklidovská vzdálenost.



Obrázek 3.2.21: Zobrazení 3 nejbližších sousedů zvoleného bodu

Nyní je potřeba získat vzdálenosti $k\_distance$ jednotlivých bodů, které odpovídají vzdálenosti bodu a jeho *3*-nejbližšího souseda. Tyto vzdálenosti jsou uvedeny v následující *Tabulce 3.2.5*.

	1	2	3	4	5	6	7	8	9	10
k-distance jednotlivých bodů	1,342	0,640	0,583	0,447	0,583	0,447	0,510	0,608	0,566	1,082

Tabulka 3.2.5: Tabulka k-distance jednotlivých bodů

Již z této tabulky je vidět, že právě první a desáté pozorování jsou vzdálenější od svého třetího souseda více nežli ostatní. K identifikaci outliers to však nestačí, a tak následujícím krokem v metodě LOF bude výpočet tzv. dosažitelné vzdálenosti $reachability\_distance_k(A,B)$ dle *vzorce (7).* Jednoduše řečeno, každé dvojici bodů bude přiřazena dosažitelná vzdálenost tak, že budou opět vypočteny Euklidovské vzdálenosti mezi jednotlivými body a následně porovnány dle *vzorce (7)* s příslušnou $k\_distance$ vzdáleností. Jako $reachability\_distance_k(A,B)$ pak bude určena ta vyšší hodnota. Přesné hodnoty dosažitelných vzdáleností lze najít v *Příloze 3.* této práce. Je také třeba mít na paměti, že tyto vzdálenosti nejsou symetrické, a tak platí, že $reachability\_distance_k(A,B) \neq reachability\_distance_k(B,A)$.

Dalším krokem metody LOF je výpočet tzv. místní dosažitelné hustoty (*lrd*), která vyžaduje dosažitelné vzdálenosti všech sousedů zvoleného bodu. Hodnoty *lrd* vypočtené dle *vzorce (8)* budou uvedeny v následující *Tabulce 3.2.6*, přičemž $|N_k(A)|$ je rovna číslu 3, neboť byli hledáni tři sousedé každého bodu.

	1	2	3	4	5	6	7	8	9	10
lrd	0,833	1,796	1,796	1,796	1,859	1,948	1,851	1,782	1,782	1,150

Tabulka 3.2.6: Tabulka místních dosažitelných hustot jednotlivých bodů

Z *Tabulky 3.2.6* však zatím o odlehlosti pozorování nelze nic určitého říci. Je tak nutné provést poslední krok metody, a to za pomoci místních dosažitelných hustot vypočítat konečný místní faktor odlehlosti (*LOF*) dle *vzorce (9).*

	1	2	3	4	5	6	7	8	9	10
LOF	2,155	1,028	1,028	1,028	1,003	0,942	0,993	1,044	1,044	1,569

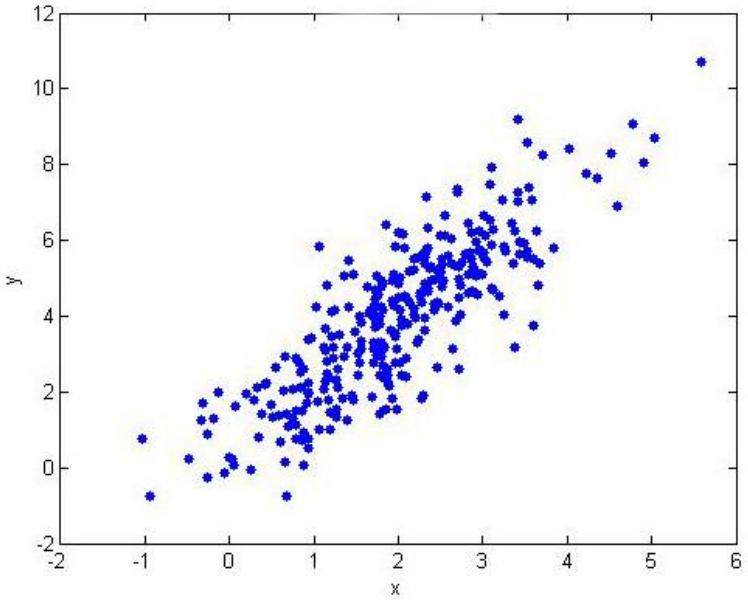
Tabulka 3.2.7: Výsledné hodnoty místního faktoru odlehlosti (LOF)

Jak již bylo zmíněno v teoretické části práce, pokud je hodnota *LOF* blízká číslu 1, lze říci, že místní dosažitelné hustoty bodu a jeho sousedů je srovnatelné a pozorování tak není odlehlé. Zatímco převyšuje-li *LOF* hodnotu 1, lze detekovat outlier, neboť číslo převyšující jedničku je způsobeno vyšší hustotou sousedů oproti hustotě zkoumaného bodu. Zde by tak mohla být identifikována pozorování 1 a 10 jako odlehlá. Tento závěr by mohl odpovídat také vizuálnímu rozložení bodů, které bylo zobrazeno například v *Obrázku 3.2.20*.

Nyní bude proveden experiment s větším počtem dat, což je reprezentováno 300 pozorováními. Jedná se o náhodně generovaná data z dvourozměrného normálního rozdělení, opět s parametry

$$\boldsymbol{\mu} = (2,4)^T \quad , \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1{,}6 \\ 1{,}6 & 4 \end{pmatrix},$$

které lze zobrazit v následujícím *Obrázku 3.2.22*.



Obrázek 3.2.22: Větší počet dat pro identifikaci outliers metodou LOF

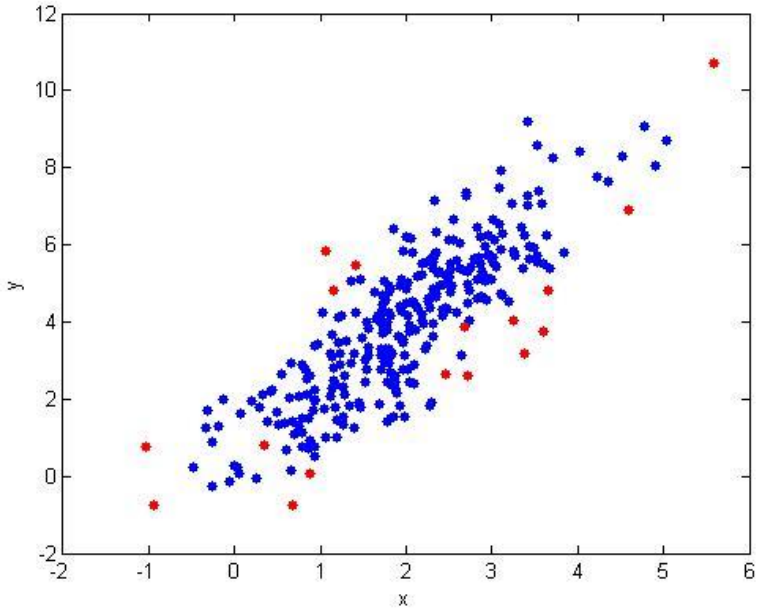
S větším počtem dat bude navýšen také počet sousedů hledaného bodu. Místo tří nejbližších sousedů jich nyní bude hledáno pět. Ve všech dalších výpočtech tak bude $k = 5$. Po aplikaci nezbytných kroků metody LOF, výpočtu $k\_distance$ vzdáleností, které nyní odpovídají vzdálenosti zkoumaného bodu a jeho pátého souseda, či výpočtu dosažitelných vzdáleností a místních dosažitelných hustot budou získány také hodnoty místního faktoru odlehlosti každého pozorování. Tyto hodnoty je možné nalézt v *Příloze 4.* této práce.

Jako outliers lze však hodnotit pouze ta pozorování, jejichž LOF převyšuje číslo 1. Jako hranici identifikace outliers byla zvolena hodnota 1,5. Ovšem záleží pouze na pozorovateli, jak hranici stanoví. Tato podmínka v tomto případě znamená, že bylo nalezeno 17 potenciálně odlehlých hodnot zobrazených níže spolu s jejich *LOF* hodnotami.

Pozorování	**9**	**15**	**21**	**35**	**39**	**54**	**94**	**162**	**165**
LOF	2,399	1,917	1,545	2,247	1,503	1,581	1,812	1,524	1,992
Pozorování	**198**	**208**	**222**	**238**	**262**	**263**	**277**	**284**	
LOF	1,737	1,832	1,959	1,569	1,508	1,663	1,711	1,773	

Tabulka 3.2.8: Outliers spolu s jejich LOF hodnotami převyšujícími zvolenou hranici

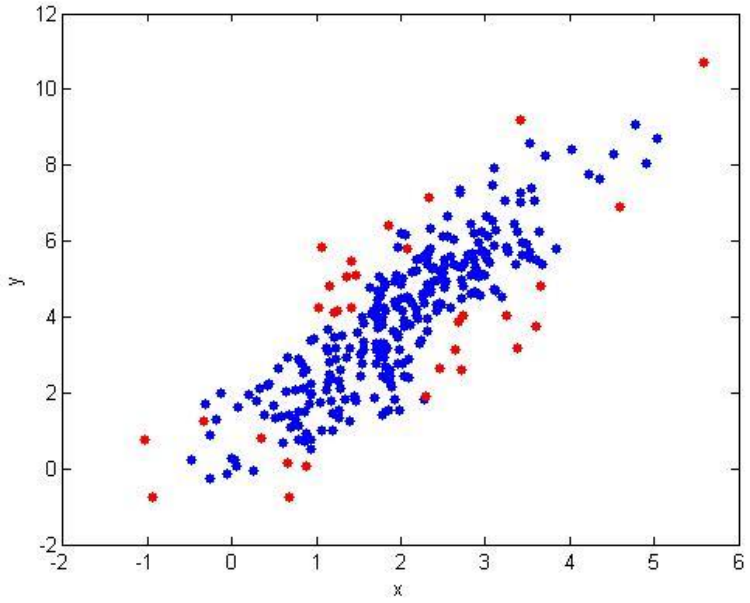
Graficky znázorněné generované hodnoty spolu s hodnotami podezřelými z odlehlosti dle metody LOF je možné znázornit následovně.



Obrázek 3.2.23: Generovaná data spolu s hodnotami podezřelými z odlehlosti

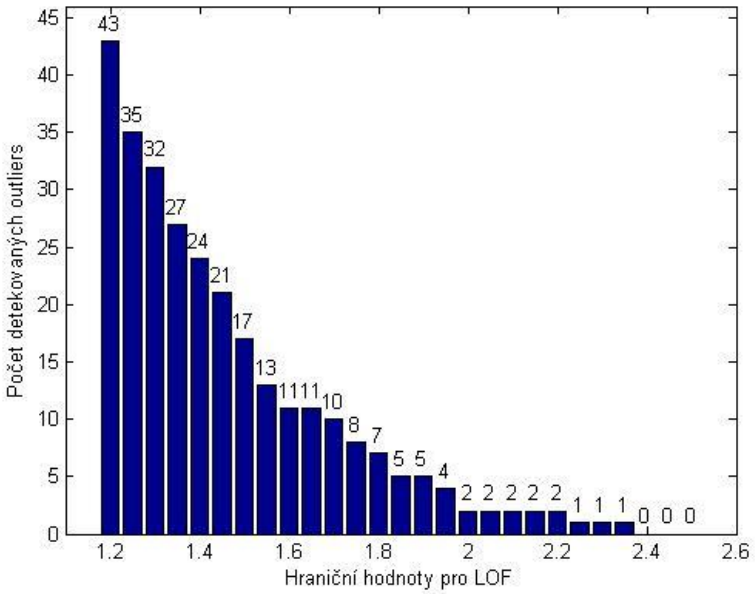
V *Obrázku 3.2.23* jsou hodnoty podezřelé z odlehlosti znázorněny červeně. Ovšem při posuzování odlehlosti hodnot velmi záleží na volbě hranice, která udává rozdíl mezi outliers a neodlehlými hodnotami. Stejně tak je ovšem důležité vyvarovat se již zmíněným efektům mnohorozměrných

dat (*Masking effect* a *Swamping effect*). Bude-li zvolena „přísnější“ hranice detekce outliers, například 1,3, bude získáno více odlehlých hodnot, a to konkrétně 32. Tato situace je názorně zobrazena v následujícím *Obrázku 3.2.24*.



Obrázek 3.2.24: Identifikované outliers dle LOF s nižší hranicí

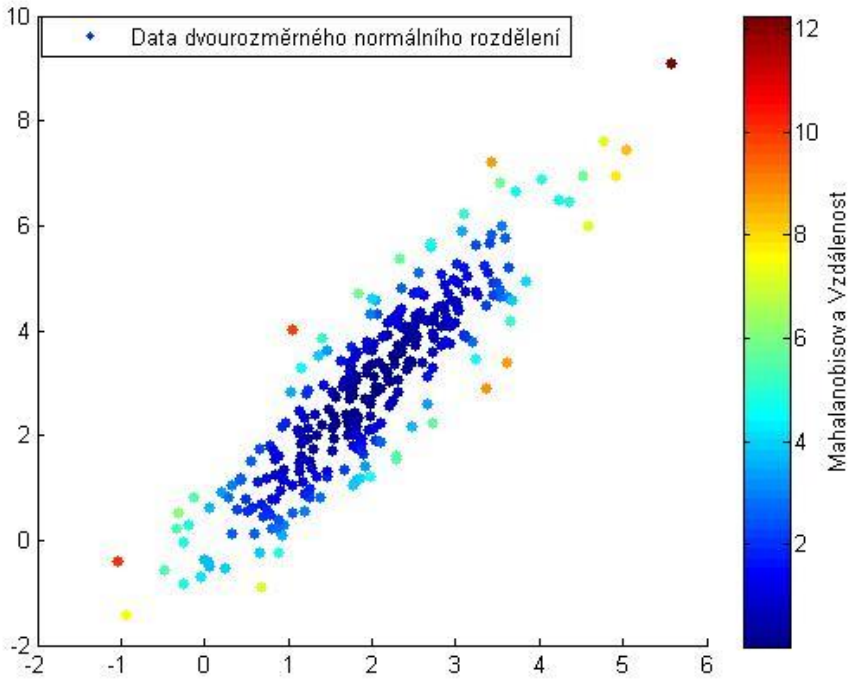
V této metodě je tedy opravdu velmi důležitá volba hraniční hodnoty pro detekci outliers, kterou pozorovatel použije. Z tohoto důvodu budou nyní v *Obrázku 3.2.8* znázorněny počty detekovaných odlehlých hodnot pro zkoumaný soubor dat a různé volby hraničních bodů metody LOF.



Obrázek 3.2.25: Počty detekovaných odlehlých hodnot při různé hranici určení outliers

Rozdíl mezi volenými hraničními body je vždy 0,05. Takto volený interval mezi jednotlivými hodnotami pro metodu LOF naznačuje exponenciální pokles počtu detekovaných odlehlých hodnot. Zatímco volba hraniční hodnoty pozorovatelem 1,2 udává počet detekovaných outliers 43 a drobná změna na hraniční hodnotu 1,25 pak o 8 odlehlých hodnot méně, volba hranice vyšší než 2 detekuje maximálně dvě odlehlé hodnoty.

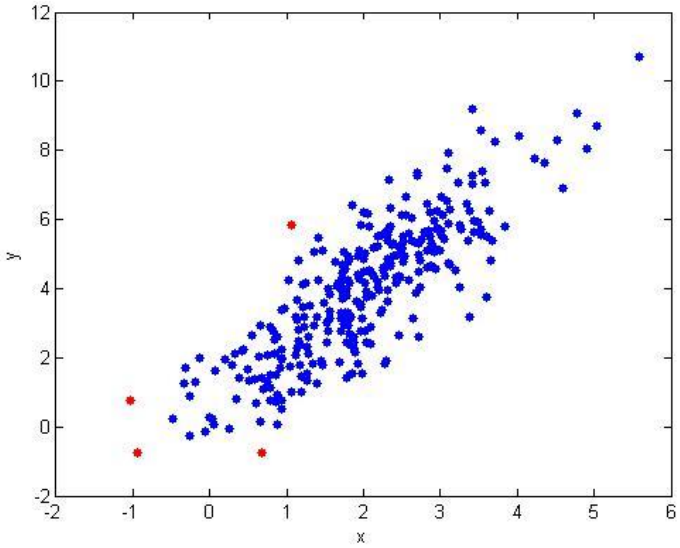
Density-based metoda LOF je velmi citlivá na volbu hraniční hodnoty detekce outliers, jak již bylo zmíněno, a také na drobné oddělení shluku dat od celého souboru, což může být vidět například také v *Obrázku 3.2.24* v levé spodní části, kde se nachází detekovaný outlier uprostřed „správných“ hodnot. Bude-li tato metoda založená na porovnání hustot sousedů jednotlivých bodů doplněna grafickou metodou Mahalanobisovy vzdálenosti definované v kapitole 2.3.1, lze vyloučit tato pozorování z množiny outliers.



Obrázek 3.2.26: Mahalanobisovy vzdálenosti v datovém souboru

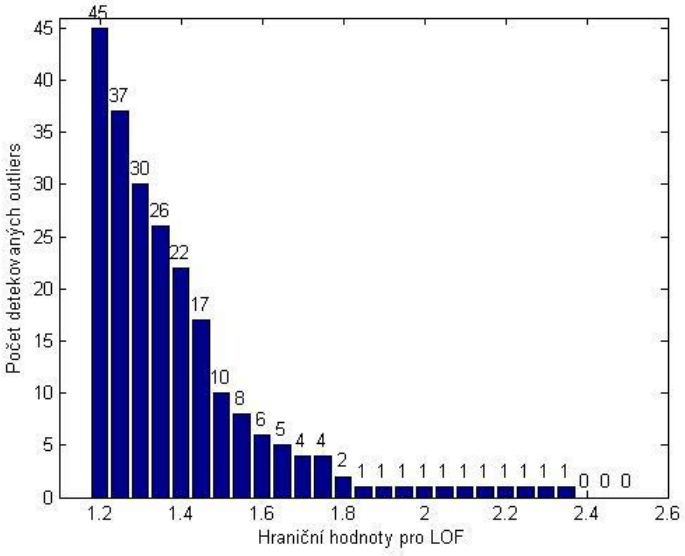
Ovšem pomocí Mahalanobisových vzdáleností bude detekováno mnohem méně odlehlých hodnot, a tak je opět na pozorovateli, aby pokud možno zkontroloval případné chyby v souboru dat a následně rozhodl, zda jsou hodnoty správné či chybné. Důležitým rozhodnutím pak zůstává, zda tyto hodnoty podezřelé z odlehlosti v souboru dat ponechá, nebo je naopak ze souboru vyřadí.

Rozdílných výsledků lze dosáhnou také v případě, že místo Euklidovské vzdálenosti při hledání *k*-nejbližších sousedů bude využita vzdálenost Mahalanobisova. Například při volbě mezní hranice detekce outliers 1,7 jsou jako odlehlé hodnoty identifikované následující, zobrazené v *Obrázku 3.2.27*.



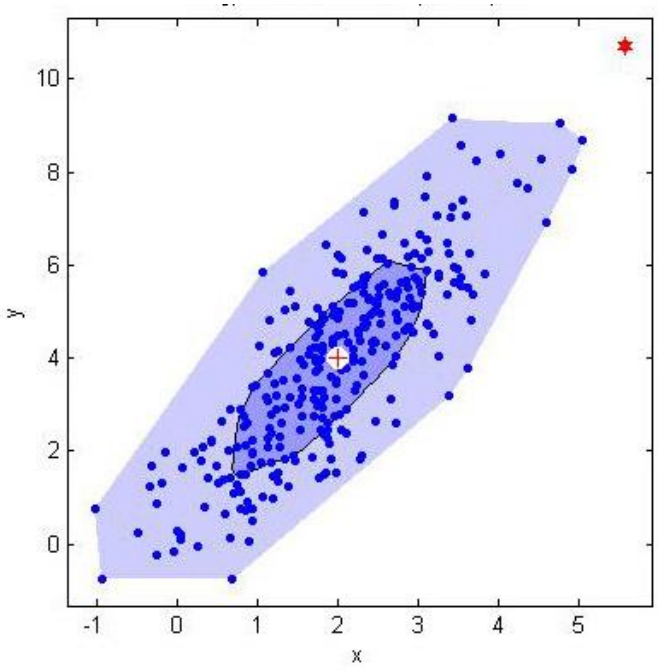
Obrázek 3.2.27: Identifikované outliers metodou LOF (Mahalanobisova vzdálenost)

Při volbě Mahalanobisovy vzdálenosti jako nástroje pro odhalení *k*-nejbližších sousedů, dochází k exponenciálnímu poklesu detekovaných odlehlých hodnot s nárůstem mezní hranice identifikace outliers rychleji než v případě Euklidovské vzdálenosti. Toto je možné doložit následující hodnotami.



Obrázek 3.2.28: Počty detekovaných outliers metodou LOF (Mahalanobisova vdálenost)

Opět je možné porovnat detekované odlehlé hodnoty metodou LOF ještě s grafickou metodou, a to bagplotem, který ovšem identifikuje pouze jediný outlier a samozřejmě nezohledňuje hustotu sousedních bodů.



Obrázek 3.2.29: Bagplot generovaných dat pro metodu LOF

Nevýhodou této metody může být opět volba hranice, kdy jsou detekovány odlehlé hodnoty, neboť drobnou změnou této hranice je detekován jiný počet outliers. Další vlastností metody je velká citlivost na efekty dat. Jen trochu větší vzdálenost bodů od většího shluku dat zapříčiňuje menší hustotu těchto bodů a následně tedy jejich detekci jako outliers.

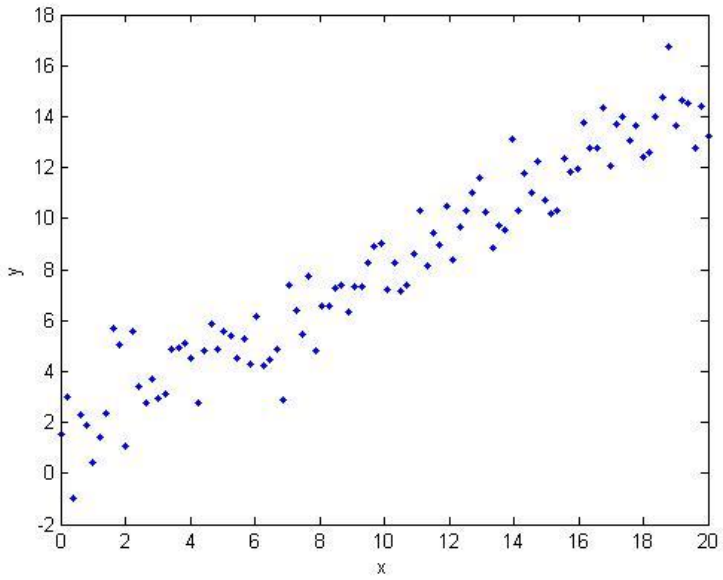
3.3 Numerické experimenty detekce vlivných bodů v regresi

Data ve formě časové řady se v praxi nejčastěji využívají k predikci budoucího vývoje zkoumaného faktoru. Zde je proto velmi důležité odhalit možné vlivné body, které by významně zkreslily výsledky dalších statistických postupů a vedly by tak k chybným predikcím budoucích hodnot.

Jak již bylo vysvětleno v teoretické části, vlivné body se v regresi rozdělují na leverage points (body odlehlé ve směru osy x) a outliers (body odlehlé ve směru osy y). Tyto odlehlé hodnoty je pak možné odhalit pomocí projekční matice H, Williamsova grafu, Cookovy vzdálenosti či Welsch-Kuhovy vzdálenosti.

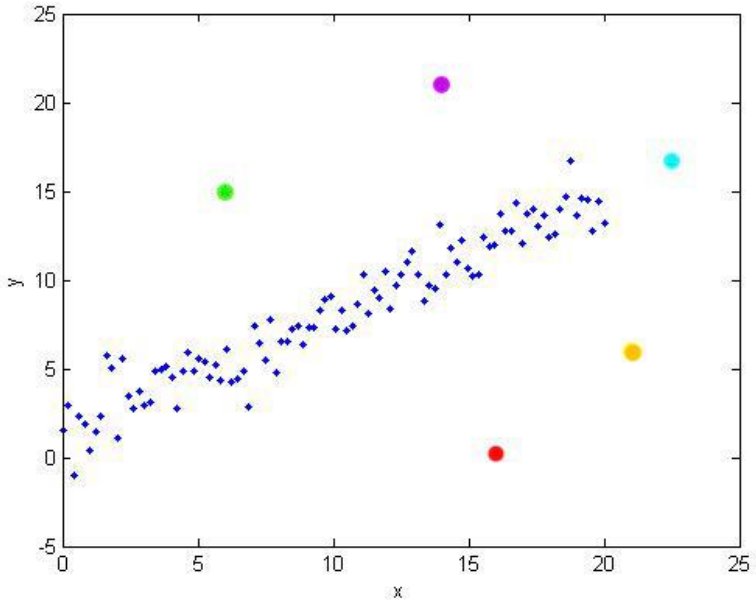
Pro účely numerických experimentů, kde budou popsány všechny tyto metody, jsou vygenerována data znázorňující časovou řadu s lineárním trendem. Tato pozorování jsou dána regresním

modelem ve tvaru $Y = \beta_0 + \beta_1 X + \varepsilon$, kde koeficienty β_0, β_1 byly zvoleny konkrétně ve tvaru $\beta_0 = 1, \beta_1 = 0{,}7$ a ε reprezentuje náhodnou chybu modelu a bude tak generováno pomocí normovaného normálního rozdělení. Pomocí takto zvolených parametrů jsou vygenerována data zobrazená na následujícím *Obrázku 3.3.1*.



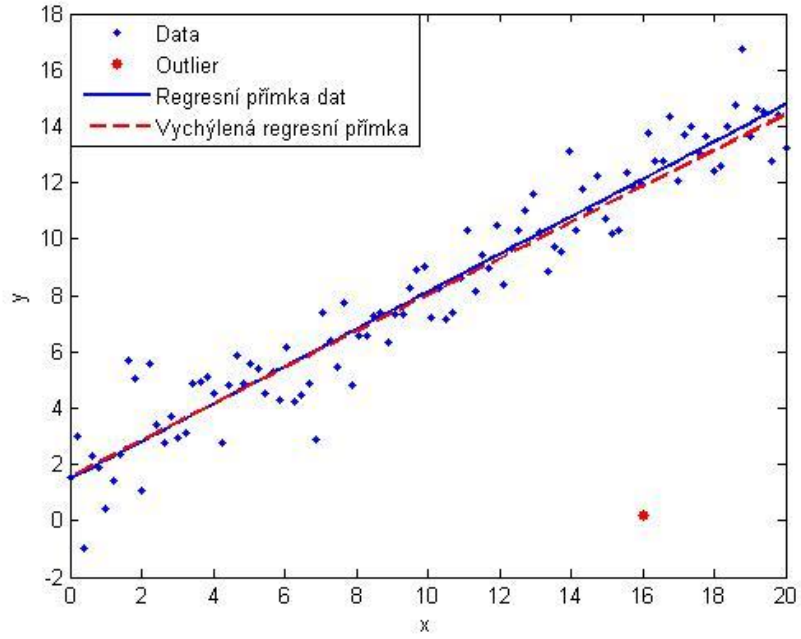
Obrázek 3.3.1: Data generovaná dle regresního modelu

Aby mohly být detekovány vlivné body, budou do tohoto souboru dat záměrně zařazeny odlehlé hodnoty, které mohou ovlivňovat regresní přímku. Zvolené vlivné body jsou zobrazeny v následujícím *Obrázku 3.3.2.*

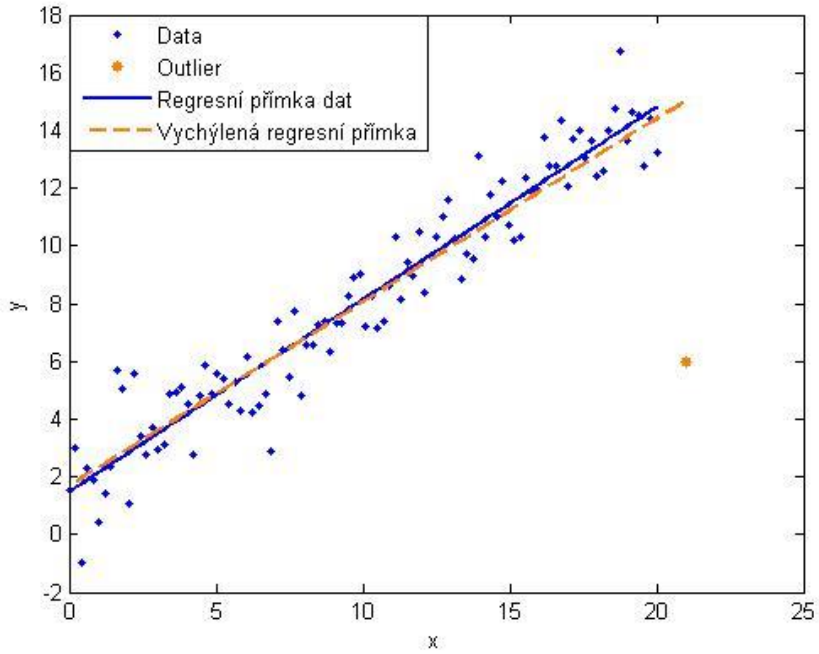


Obrázek 3.3.2: Generovaná data spolu se zvolenými vlivnými body

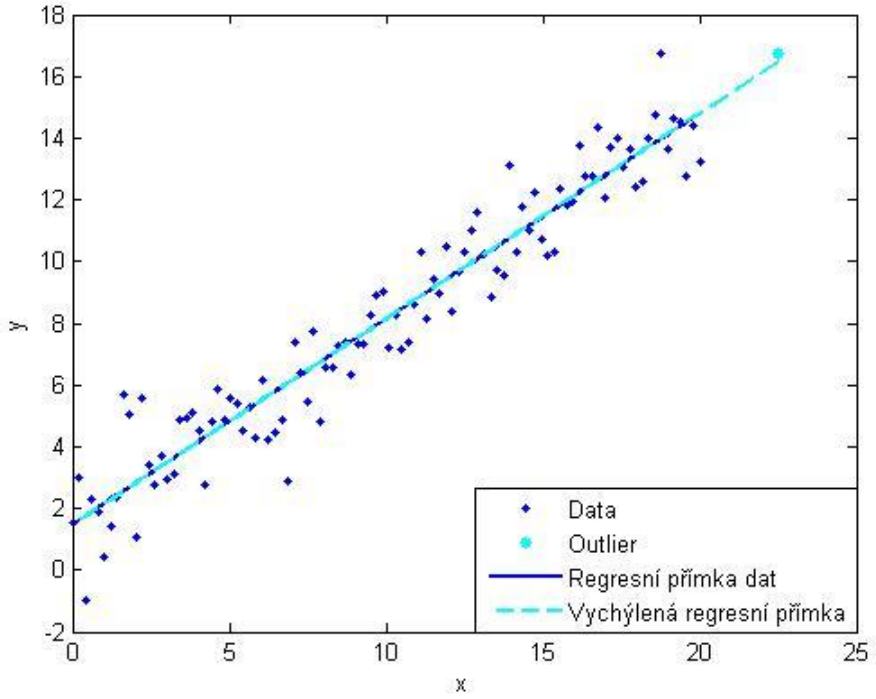
Každý jednotlivý bod, který byl zvolen jako vybočující, ovlivňuje také regresní přímku dat původních, vygenerovaných podle zvoleného regresního modelu. Tyto regresní přímky byly vypočteny pomocí programu Matlab a jeho funkcí *polyfit* a *polyval.* Znázorněny budou na následujících *Obrázcích 3.3.3, 3.3.4, 3.3.5, 3.3.6* a *3.3.7.*



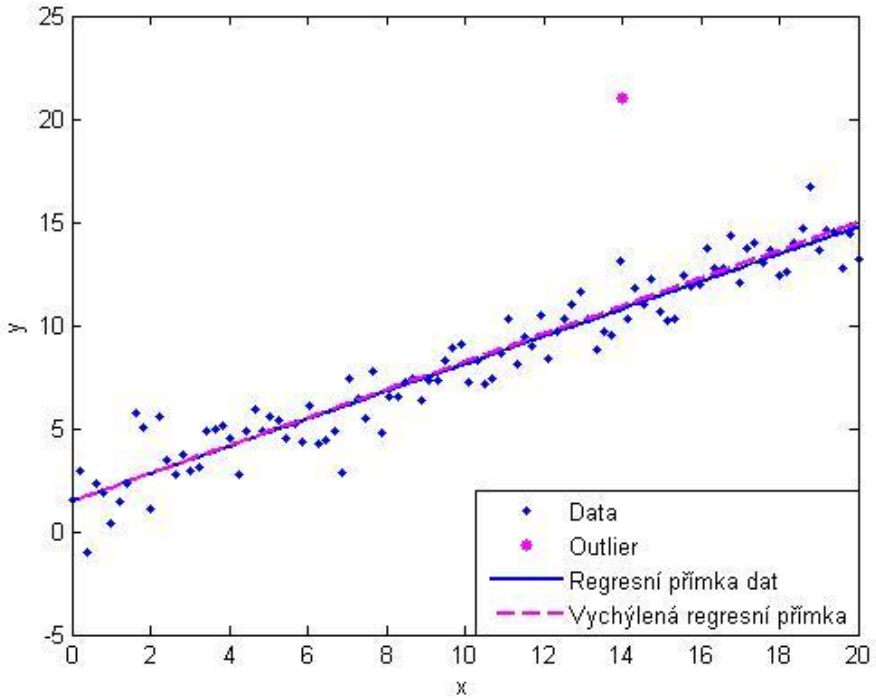
Obrázek 3.3.3: Regresní přímky dat a dat s vlivným bodem



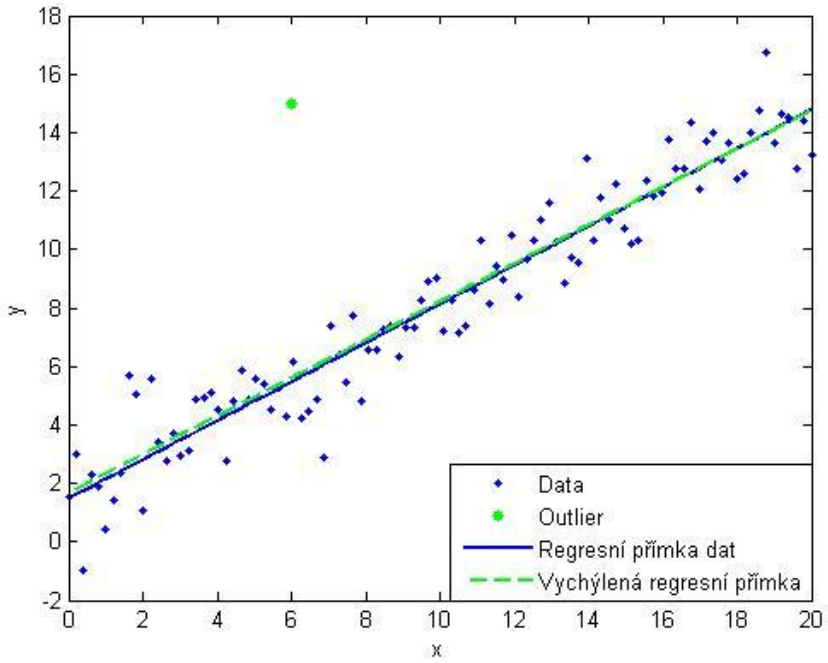
Obrázek 3.3.4:Regresní přímky dat a dat s vlivným bodem



Obrázek 3.3.5: Regresní přímky dat a dat s vlivným bodem

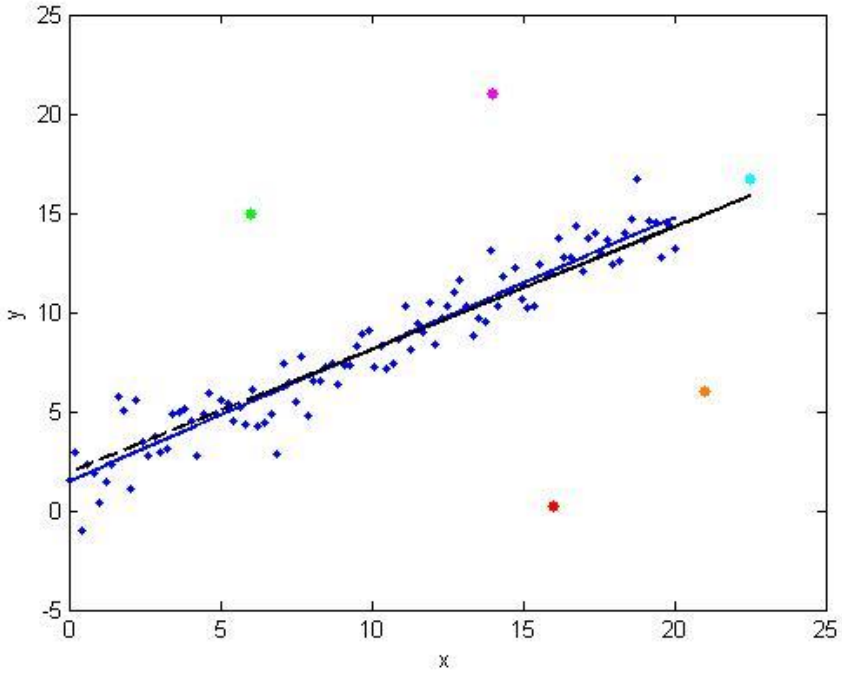


Obrázek 3.3.6: Regresní přímky dat a dat s vlivným bodem



Obrázek 3.3.7: Regresní přímky dat a dat s vlivným bodem

Následně budou zobrazeny všechny zvolené vlivné body spolu s generovanými daty a jejich regresními přímkami. Regresní přímka zohledňující všechny záměrně vlivné body je znázorněna černou barvou na *Obrázku 3.3.8.*



Obrázek 3.3.8: Všechny záměrně vlivné body a regresní přímka

Tato grafická zobrazení je vhodné doložit konkrétními změnami parametrů jednotlivých regresních přímek. Původní data byla generována dle regresního modelu $Y = 1 + 0{,}7 \cdot X + \varepsilon$, kde ε reprezentuje náhodnou chybu modelu odpovídající normovanému normálnímu rozdělení, tedy $\varepsilon \sim N(0{,}1)$ a generováno bylo celkem 100 hodnot. Koeficienty regresních přímek jsou doloženy v následující *Tabulce 3.3.1*.

		Generovaná data	Generovaná data spolu s vlivnými body					
Outlier		-	červený	oranžový	modrý	fialový	zelený	vše
Koeficienty	$\boldsymbol{\beta_0}$	**1,474**	**1,563**	**1,675**	**1,466**	**1,456**	**1,678**	**1,924**
	Absolutní rozdíl	-	0,089	0,202	0,008	0,018	0,205	0,451
	$\boldsymbol{\beta_1}$	**0,665**	**0,644**	**0,636**	**0,666**	**0,677**	**0,654**	**0,620**
	Absolutní rozdíl	-	0,021	0,029	0,001	0,012	0,011	0,045

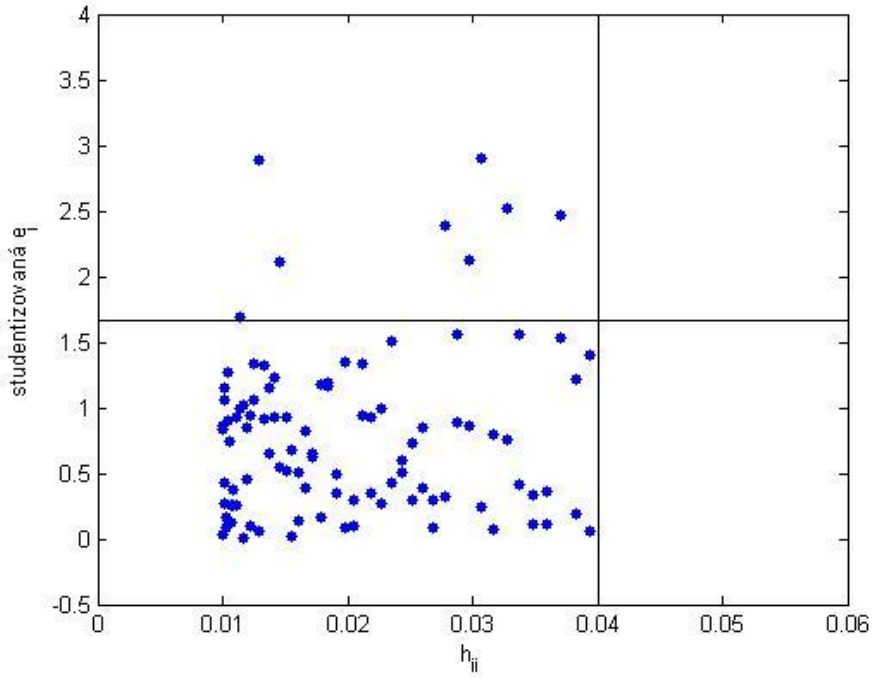
Tabulka 3.3.1: Koeficienty regresních přímek generovaných dat s různými vlivnými body

Díky koeficientům zobrazeným v *Tabulce 3.3.1* lze konstatovat, že například světle modrý bod, graficky vyobrazen také v *Obrázku 3.3.5,* není příliš vlivným bodem, neboť respektuje trend generovaných dat bez jakýchkoliv outliers. Naopak největší změna regresní přímky nastává při kombinaci všech záměrně zvolených vlivných bodů.

Toto však není žádná metoda detekce vlivných bodů, aby mohlo být řečeno, který záměrně zvolený bod mimo generovaná data je odlehlý, ať už outlier či leverage point. Právě proto budou nyní aplikovány metody detekce vlivných bodů v regresi, které byly popsané v teoretické části práce.

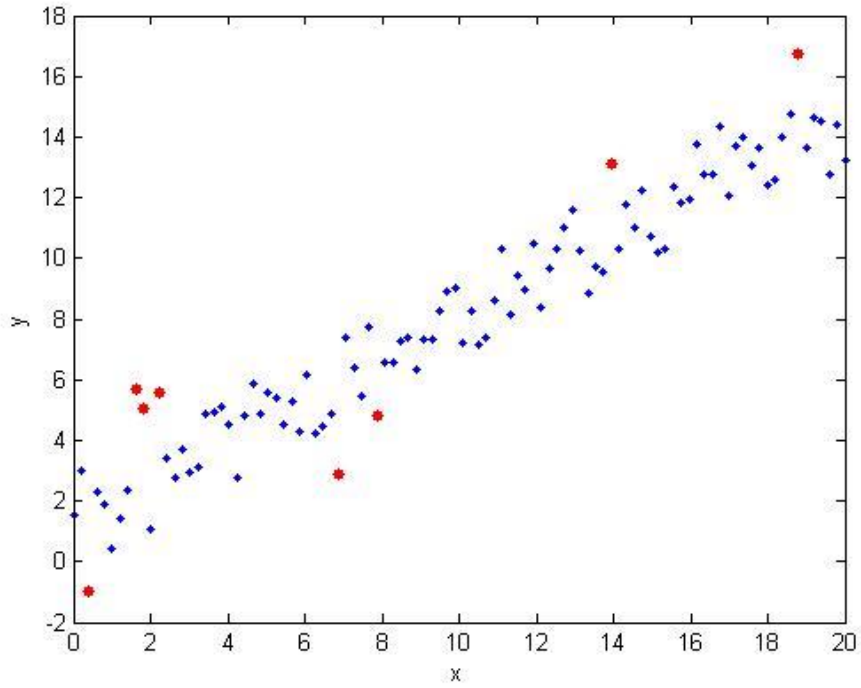
3.3.1 Williamsův graf

Tato metoda graficky zobrazuje závislost Studentizovaných jackknife reziduí na diagonálních hodnotách projekční matice H. Pro generovaná data dle zvoleného regresního modelu je Williamsův graf zobrazen v následujícím *Obrázku 3.3.9,* ze kterého lze identifikovat vlivné body ve směru osy y, ačkoli nebyly záměrně zvoleny a jsou tak výsledkem samotného generování a náhodné chyby modelu.



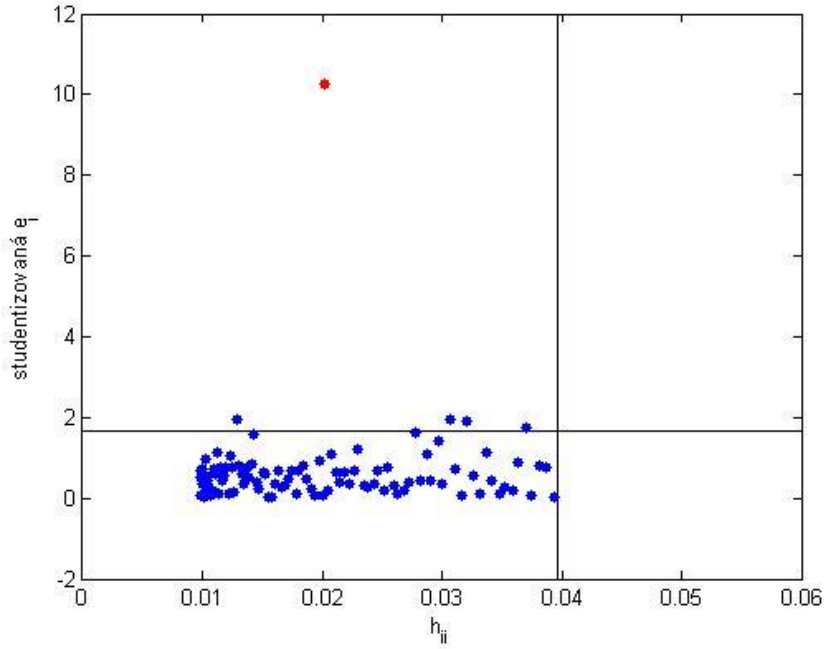
Obrázek 3.3.9: Williamsův graf generovaných hodnot dle regresního modelu

Takovýchto vlivných bodů ve směru osy y bylo v datech nalezeno pomocí Williamsova grafu hned 8. Budou-li tyto outliers zobrazeny v původním datovém souboru, bude přehledněji vidět vliv těchto bodů ve směru osy y.



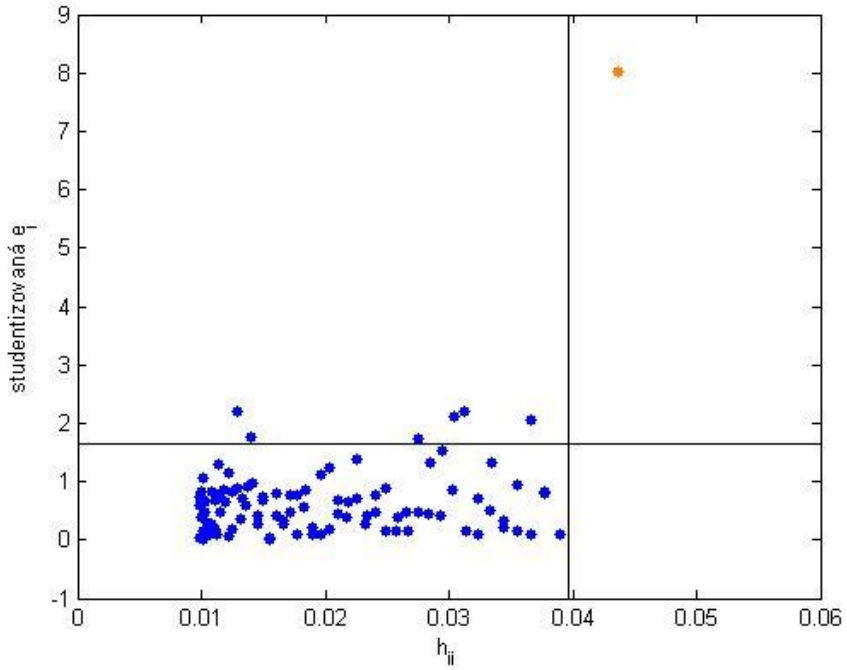
Obrázek 3.3.10: Data pro regresi s outliers podle Williamsova grafu

Budou-li dále přidány jednotlivé záměrně vlivné body, zobrazené také v *Obrázku 3.3.2*, budou získány Williamsovy grafy.



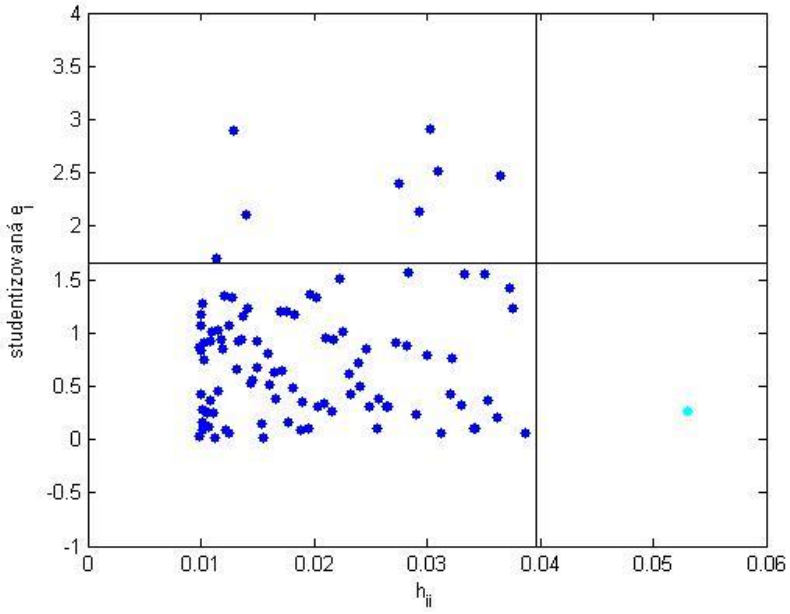
Obrázek 3.3.11: Williamsův graf s prvním zvoleným (červeným) bodem

Tímto přidaným záměrně vlivným bodem se mimo jiné také posunula mez detekce outliers, a tak byly identifikovány čtyři outliers generovaných dat a samozřejmě bod záměrně odlehlý.



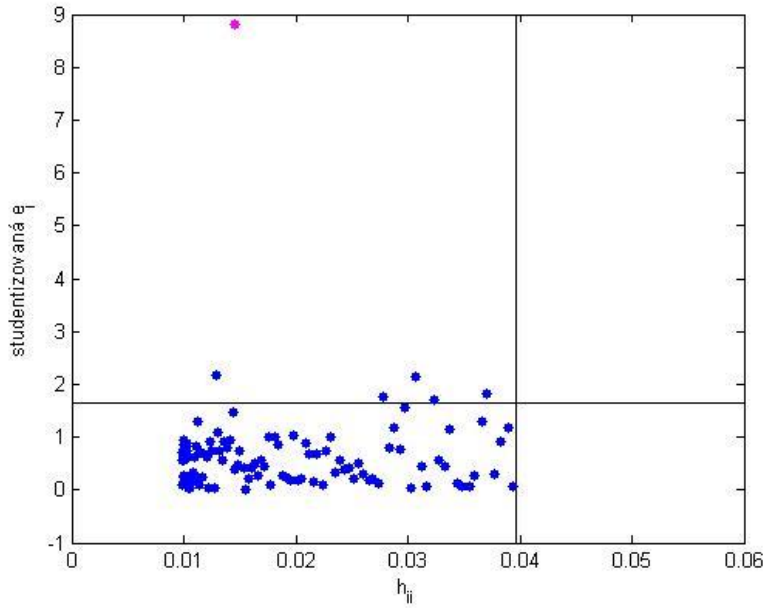
Obrázek 3.3.12: Williamsův graf s druhým zvoleným (oranžovým) bodem

Z *Obrázku 3.3.12* je patrné, že druhý zvolený bod je vlivným bodem ve směru osy x a zároveň také ve směru osy y. Toto lze pozorovat také na *Obrázku 3.3.4,* kde je tento bod zobrazen s celým souborem generovaných hodnot.

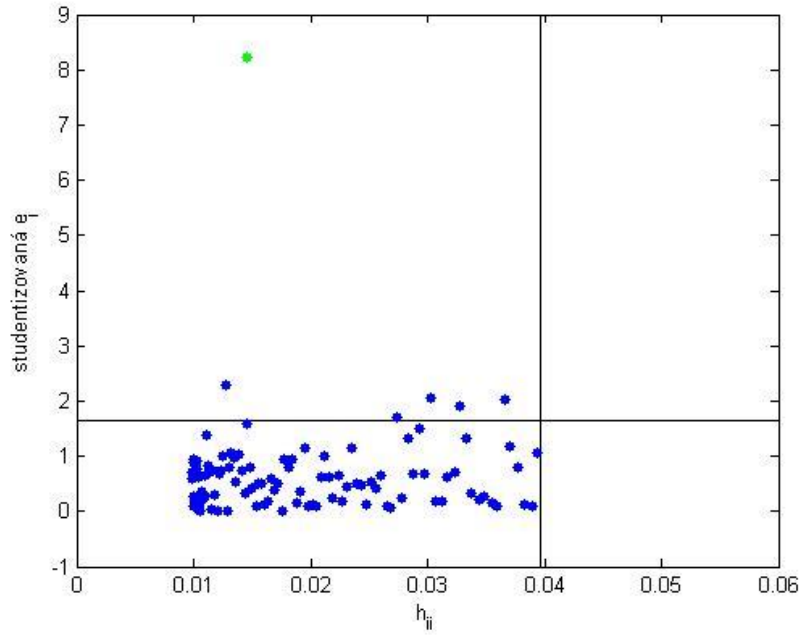


Obrázek 3.3.13: Williamsův graf se třetím zvoleným (světle modrým) bodem

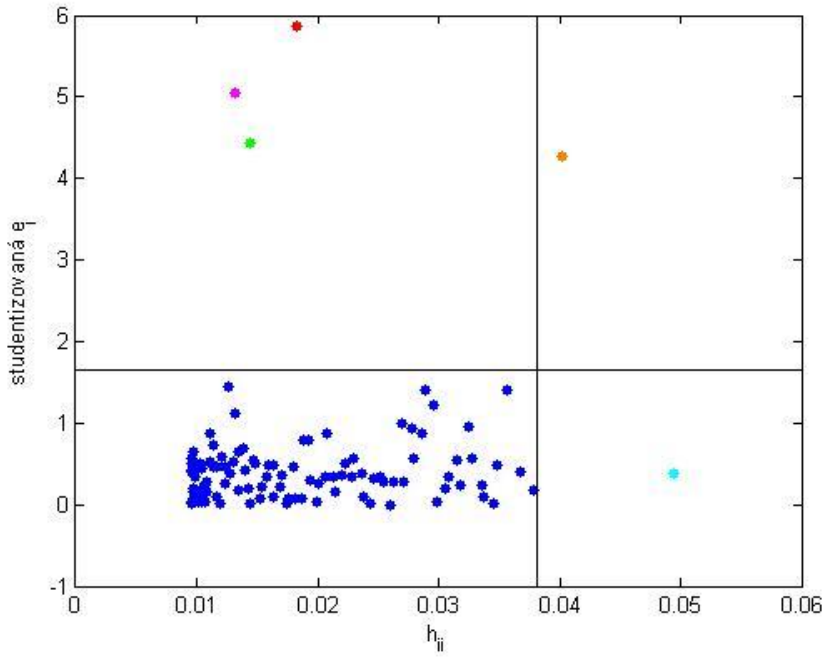
Tento bod je charakteristický tím, že respektuje trend zvoleného regresního modelu, a tak není detekován jako outlier, ale pouze jako leverage point, tedy vybočující ve směru osy x od zbylých dat.



Obrázek 3.3.14: Williamsův graf se čtvrtým zvoleným (fialovým) bodem



Obrázek 3.3.15: Williamsův graf s pátým zvoleným (zeleným) bodem



Obrázek 3.3.16: Williamsův graf se všemi zvolenými body

Díky *Obrázku 3.3.16* lze identifikovat všech 5 záměrně zvolených vlivných bodů, konkrétně pak tři outliers, jeden leverage point a jeden outlier a zároveň leverage point. Toto tvrzení je možné doložit také *Obrázkem 3.3.2,* kde jsou zobrazeny vlivné body, a to dva odlehlé body ve směru osy x a čtyři odlehlé body ve směru osy y.

3.3.2 Detekce leverage points pomocí projekční matice H

Nyní půjde o to, detekovat pouze leverage points, tedy body odlehlé ve směru osy x. Z předešlých Williamsových grafů aplikovaných na tatáž data lze předpokládat, že budou odhaleny dva leverage points, které byly zbarveny světle modrou a oranžovou barvou.

Leverage points jsou pomocí této metody detekovány díky projekční matici $H = X(X^T X)^{-1}X^T$ a jejích diagonálních prvků. V programu Matlab lze tyto hodnoty získat příkazem *regstats,* kde se právě diagonální prvky projekční matice skrývají pod pojmenováním *leverage.*

Tato metoda detekce vlivných bodů ve směru osy x udává, že za leverage point lze považovat takový bod, jehož hodnota diagonálního prvku v projekční matici převyšuje poměr stopy matice H vynásobený dvěma a počtu hodnot v souboru dat, tedy pokud platí $h_{ii} > \frac{2p}{n}$. Tímto způsobem byly detekovány dva leverage points přesně dle předpokladů, a to druhý (oranžový) a třetí (světle modrý) bod. Toto tvrzení lze doložit konkrétními hodnotami uvedenými v následující *Tabulce 3.3.2.*

	Generovaná data spolu se všemi vlivnými body				
Outlier	červený	oranžový	modrý	fialový	zelený
h_{ii}	0,018	**0,040**	**0,049**	0,013	0,014
$2p/n$	0,038	0,038	0,038	0,038	0,038
Leverage point	**-**	**ano**	**ano**	**-**	**-**

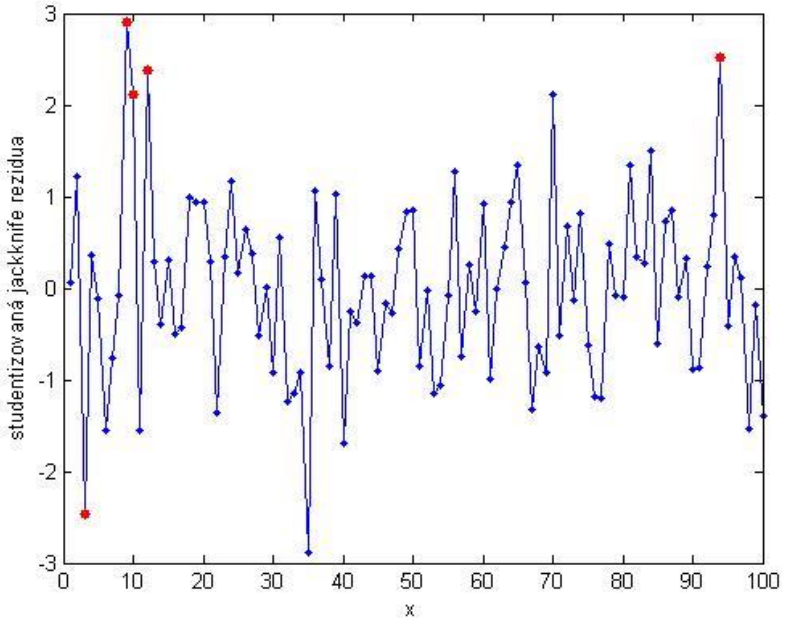
Tabulka 3.3.2: Hodnoty detekující leverage points

3.3.3 Cookova vzdálenost a Welsch-Kuhova vzdálenost

Obě tyto vzdálenosti pomáhají pozorovateli detekovat outliers, tedy vlivné body ve směru osy y. Stejně tak jsou obě metody založené na Studentizovaných jackknife reziduích. Rozdílem je, že Cookova vzdálenost měří vliv i-tého pozorování na hodnotu odhadu vektoru β regresního modelu, zatímco Welsch-Kuhova vzdálenost měří vliv i-tého pozorování nejen na hodnotu odhadu vektoru β, ale simultánně také na odhad parametru σ^2. V programu Matlab lze opět tyto hodnoty získat příkazem *regstats,* kde se právě Cookova vzdálenost i-tého pozorování skrývá pod pojmenováním *cookd* a Welsch-Kuhova vzdálenost i-tého pozorování pod pojmenováním *dffits.*

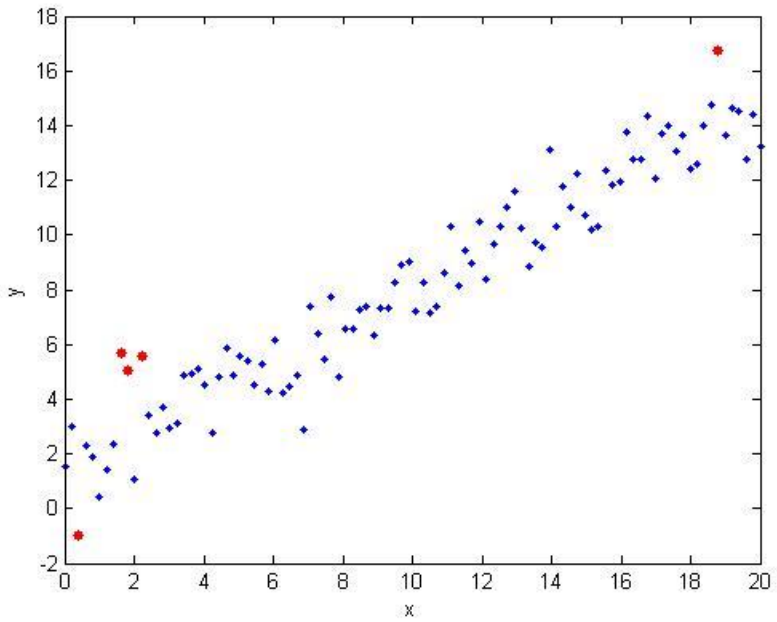
Jelikož jsou obě metody založené na Studentizovaných jackknife reziduích a jejich velikostech, budou na následujícím *Obrázku 3.3.17* zobrazena právě Studentizovaná rezidua vygenerovaného souboru dat pomocí zvoleného regresního modelu. Dále jsou pak červeně označena rezidua, která byla pomocí Cookovy vzdálenosti detekována jako rezidua hodnot podezřelých z odlehlosti. Tedy

taková rezidua, jejichž pomocí byla vypočtena hodnota Cookovy vzdálenosti větší než kvantil $F_\alpha(p, n-p)$. V tomto konkrétním případě byly vzdálenosti porovnávány s kvantilem $F_{0,05}(2; 98) = 0{,}0513$. Tuto mez v základním generovaném souboru překročilo pět Cookových vzdáleností.



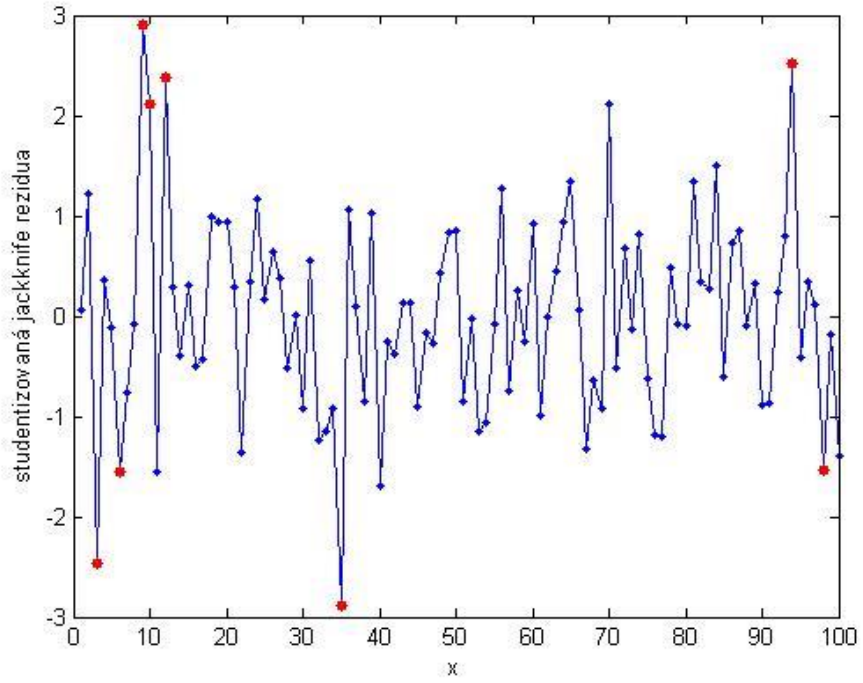
Obrázek 3.3.17: Studentizovaná jackknife rezidua generovaného souboru dat (Cookova vzdálenost)

Dle Cookovy vzdálenosti jsou tak jako outliers podezřelé následující body.

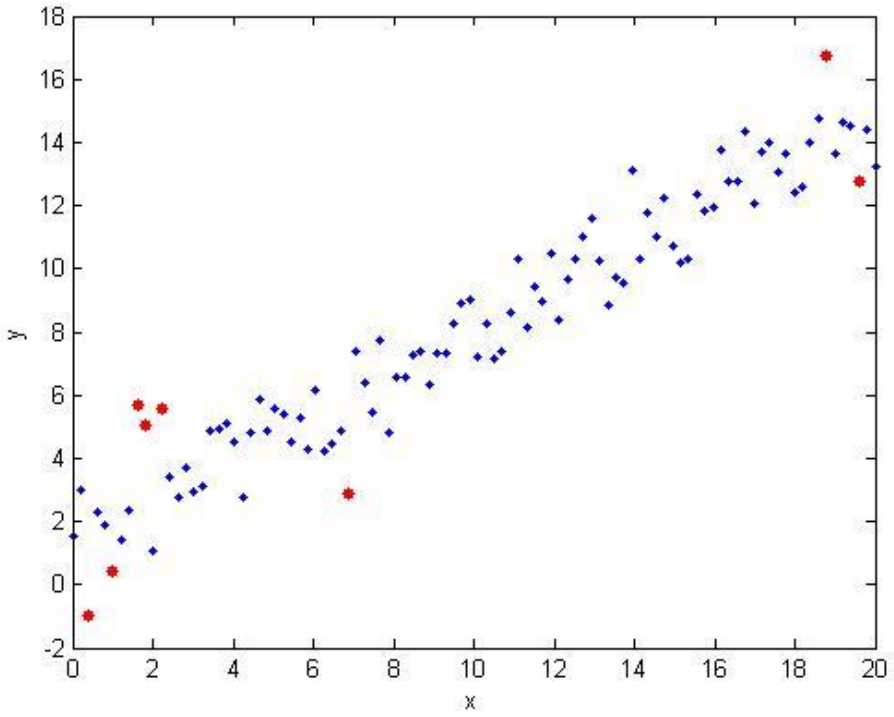


Obrázek 3.3.18: Detekované outliers pomocí Cookovy vzdálenosti

Bude-li na stejných datech využita místo Cookovy vzdálenosti Welsch-Kuhova vzdálenost, odhalí se více outliers, a to stejné body jako v předešlém případě plus další tři.



Obrázek 3.3.19: Studentizovaná jackknife rezidua generovaného souboru dat (Welsch-Kuhova vzdálenost)



Obrázek 3.3.20: Detekované outliers pomocí Welsch-Kuhovy vzdálenosti

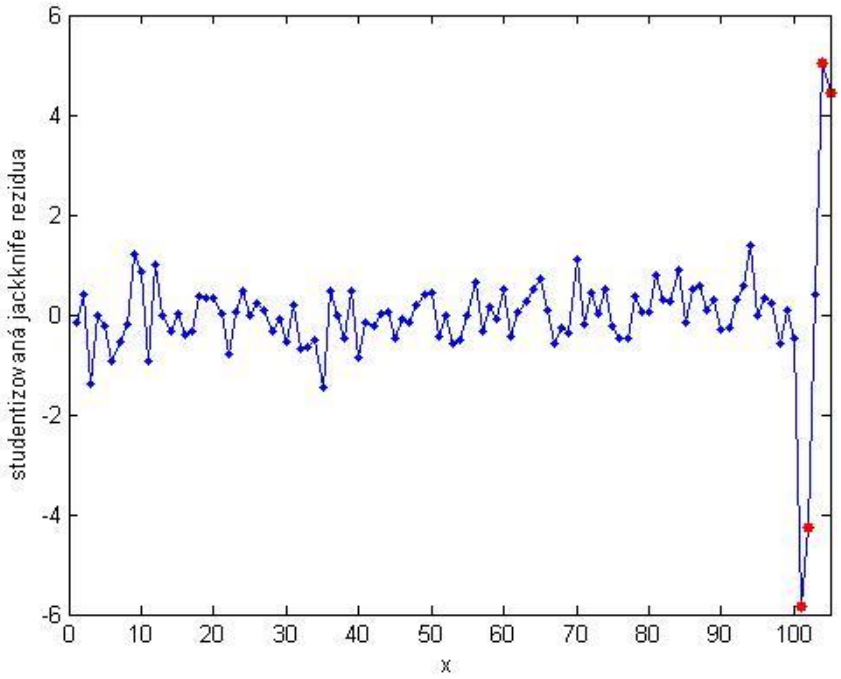
Pomocí obou vzdáleností, Cookovy i Welsch-Kuhovy, byly dále detekovány outliers v souborech dat vždy s jedním záměrně vloženým vlivným bodem a na závěr také v kombinaci všech takto zvolených pozorování spolu s generovanými daty. Získané výsledky jsou zapsány v následující *Tabulce 3.3.3*, kde čísla detekovaných outliers určují pořadí vlivného bodu v souboru dat. Přičemž v základním souboru se nachází 100 pozorování. Dále je vždy záměrně vložen jeden vlivný bod a na závěr jsou využity všechny tyto body, které je možné názorně vidět opět v *Obrázku 3.3.2.*

	Generovaná data spolu s vlivnými body						
	-	červený	oranžový	modrý	fialový	zelený	vše
Cookova vzdálenost	3;9;10; 12;94	3;9;94; 101	3;9;94;101	3;9;10; 12;94	3;9;101	3;9;94; 101	101;102; 104;105
Welsch-Kuhova vzdálenost	3;6;9;10; 12;35;94; 98	3;9;94; 101	3;9;12;94; 101	3;6;9;10; 12;35;94; 98	3;9;12; 94;101	3;9;94; 101	101;102; 104;105

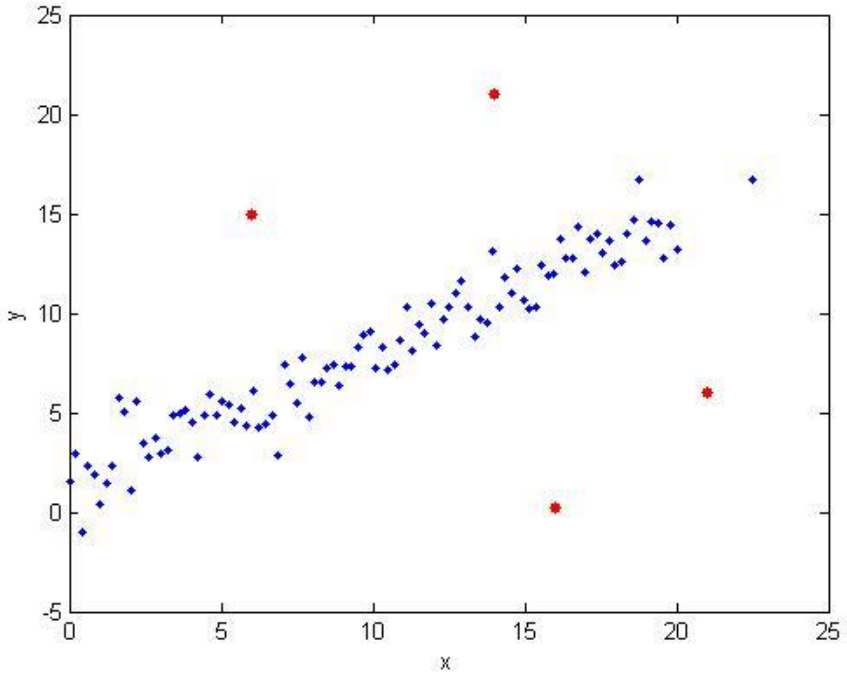
Tabulka 3.3.3: Detekové outliers pomocí Cookovy a Welsch-Kuhovy vzdálenosti

Z těchto hodnot je patrné, že obě metody detekce vlivných bodů odhalily podobná pozorování. Pomocí Cookovy vzdálenosti bylo sice detekováno méně bodů, ale pro obě metody platí, že v případě souboru generovaných dat bez přidání jakéhokoliv záměrně vlivného bodu dochází k odhalení stejných pozorování jako v případě, kdy je k souboru připojen světle modrý bod, který respektuje trend zvoleného regresního modelu a není tak vlivným bodem, ale pouze vybočujícím. V dalších čtyřech možnostech, kdy byl k základnímu souboru dat přidělen vždy jeden záměrně vlivný bod, došlo k identifikaci taktéž právě tohoto bodu, a to pomocí obou metod.

V případě generovaných dat spolu se všemi záměrně zvolenými body, byly pomocí obou metod detekovány právě čtyři outliers. Mezi záměrně zvolenými body chybí pozorování v pořadí 103, které však respektuje trend regresního modelu, a tak není identifikováno jako vlivný bod. Toto tvrzení lze opět doložit grafy Studentizovaných jackknife reziduí spolu s označenými rezidui hodnot, které jsou podezřelé z odlehlosti.



Obrázek 3.3.21: Studentizovaná jackknife rezidua generovaného souboru dat se všemi záměrně zvolenými vlivnými body



Obrázek 3.3.22: Detekované outliers celého souboru dat se všemi záměrně zvolenými vlivnými body

4 Reálná data

Detekce odlehlých pozorování má široké využití i v praxi, kdy je nutné hledat z nějakého důvodu vybočující hodnoty od běžných a rozhodnout, zda jsou to hodnoty chybné nebo naopak extrémní, které mohou udávat změnu zkoumaného faktoru.

Často se detekce odlehlých hodnot využívá také v pojišťovnictví, bankovnictví nebo lékařství či jiných oborech, kde pomáhá odhalovat podvodná jednání klientů nebo odlišné reakce na léčbu.

Pro účely této diplomové práce byla využita data, která udávají poruchovost strojů v závislosti na vlhkosti nebo teplotě v průběhu doby zkušebního provozu. Získána byla z výsledné práce smluvního výzkumu NTIS za rok 2014, který je veden pod číslem zakázky 52 9181. Tento smluvní výzkum byl realizován na katedře matematiky, Fakulty aplikovaných věd na Západočeské univerzitě. K dispozici je 4861 záznamů, které obsahují čas kontroly současně s naměřenou teplotou a vlhkostí. Následně je každá tato kontrola doplněna informací, zda byla zjištěna porucha stroje, případně jakého druhu porucha byla. Poruchy jsou označeny následujícím způsobem:

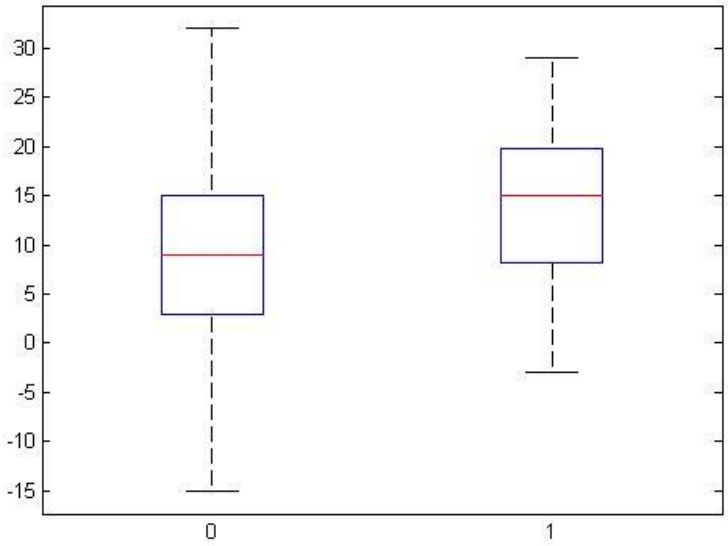
- Ok – při kontrole nebyla nalezena žádná porucha (označení 0[1])
- Skiip – technologie zaručující nízký tepelný odpor a stálost při tepelném cyklování (označení 1)
- Řídící karta (označení 2)
- Trakční motor (označení 3)
- Čerpadlo (označení 4)
- Odporník (označení 5)
- Pomocný pohon (označení 6)

Bez jakékoliv poruchy se v souboru dat nachází celkem 4774 záznamů, zatímco s poruchami pak 87. Detekce odlehlých hodnot může být uplatněna jak pro jednorozměrná data teplot (pro hodnoty bez poruch a hodnoty s poruchami) a vlhkosti, tak pro vícerozměrná data poruchovosti při závislosti na obou zvolených parametrech, tedy vlhkosti a teplotě.

Jako první budou zkoumány data pomocí jednorozměrných metod detekce outliers. Jelikož hodnoty jednotlivých skupin neodpovídají normálnímu rozdělení, nelze využít metody typu

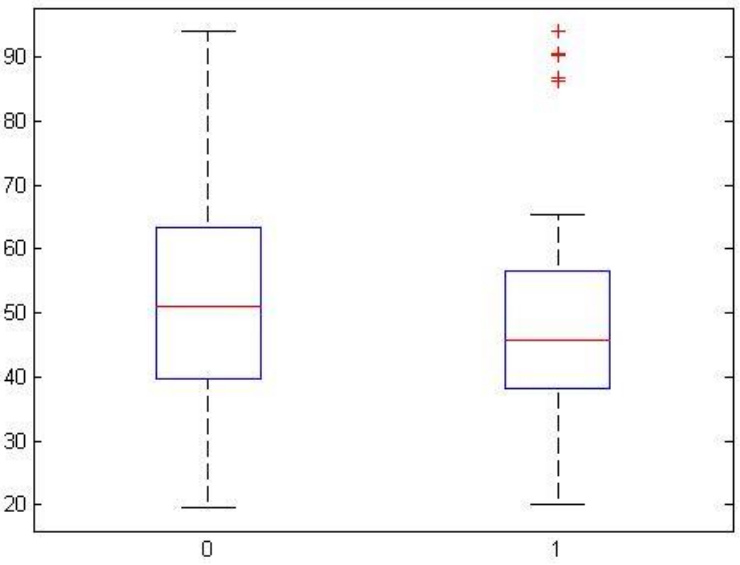
[1] Označení uvedené vždy za konkrétním typem poruchy slouží k orientaci ve výpočtech v programu Matlab.

Grubbsův test či Dean-Dixonův test, které právě normální rozdělení předpokládají. Z tohoto důvodu budou využity pouze boxploty. Získané budou následující grafy.



Obrázek 4.1: Boxploty podle teploty

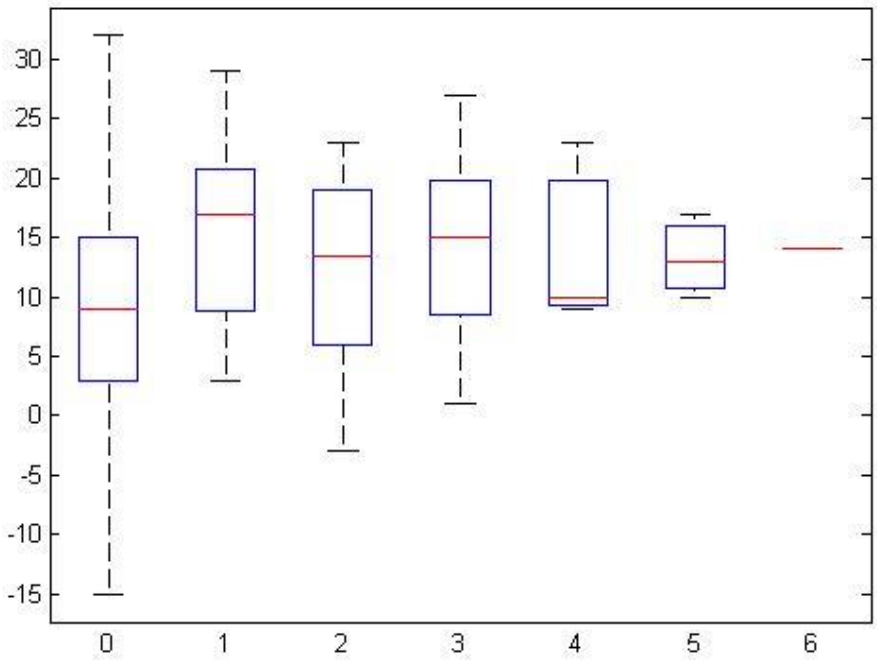
Obrázek 4.1 znázorňuje boxploty teploty dat bez poruch (označené 0) a dat, kde se vyskytly poruchy (označené 1). Díky těmto boxplotům nebyly nalezeny žádné odlehlé hodnoty teplot jednotlivých skupin dat (bezporuchová a poruchová).



Obrázek 4.2: Boxloty podle vlhkosti

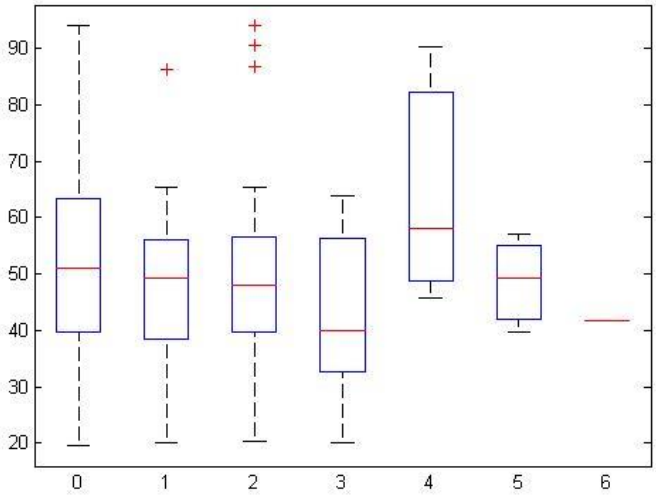
Naopak v *Obrázku 4.2,* který uvádí boxploty vlhkosti, bylo nalezeno několik outliers ve skupině dat s poruchami. Tyto odlehlé hodnoty jsou zobrazeny jako červené křížky. Z těchto boxplotů tak lze vyčíst, že v případě poruch byly naměřeny vlhkosti extrémně vysoké (kolem 90 %), které jsou však odlehlé od většího počtu hodnot, neboť mediánem vlhkosti ve skupině poruch je 45,67 %.

Dále se nabízí, prozkoumat jednotlivé typy poruch z hlediska teploty a vlhkosti.



Obrázek 4.3: Boxploty teploty podle typu poruchy

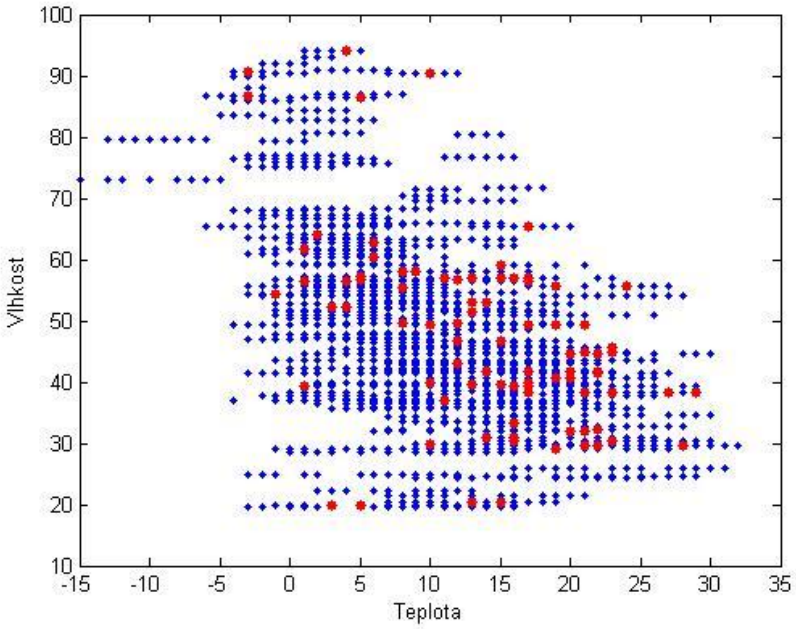
V *Obrázku 4.3* jsou vyobrazeny boxploty dat podle typu poruchy spolu s daty bez poruch (označené 0). Pod označením 6 se skrývá porucha pomocného pohonu. K této poruše došlo pouze v jediném případě, a tak je „boxplot“ uveden pouze jako mediánová linie. Opět zde není nalezen žádný outlier, a tak budou nyní zkoumána data podle vlhkosti, kde jsou nějaké odlehlé hodnoty očekávané již díky *Obrázku 4.2*. Jen zatím nebylo specifikováno, o jaký typ poruchy se jedná. Tomuto by měl být nápomocný následující *Obrázek 4.4.*



Obrázek 4.4: Boxploty vlhkosti podle typu poruchy

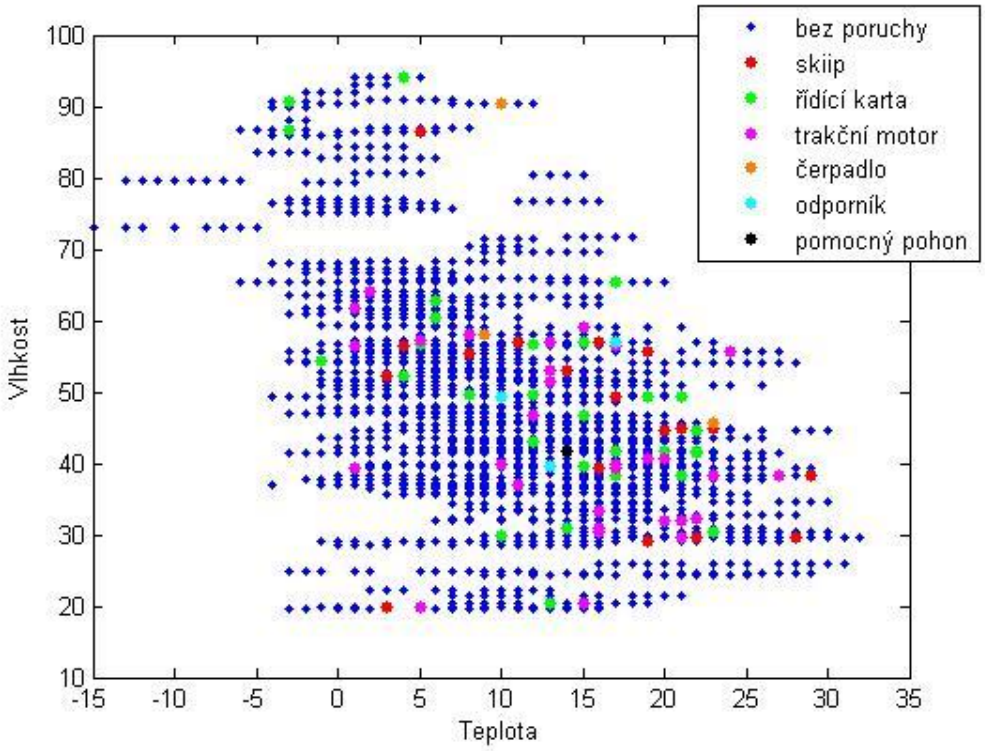
V *Obrázku 4.4* jsou blíže znázorněny nalezené odlehlé hodnoty vlhkosti při poruchách. Jeden outlier se vyskytl při poruše skiip (označené 1) a dále 3 outliers při poruše řídící karty (označené 2). V boxplotech je opět znázorněno, že mediánová hodnota těchto poruch se pohybuje mírně pod 50 %, zatímco outliers se vyskytují kolem 90 % vlhkosti.

Tato data lze však zkoumat také z pohledu vícerozměrných metod. Budou-li současně zkoumány teploty a vlhkosti jednotlivých kontrol, bude získán následující *Obrázek 4.5,* kde jsou modrou barvou zobrazeny bezporuchové kombinace teplot a vlhkostí a červenou pak poruchové.



Obrázek 4.5: Zobrazení vícerozměrných reálných dat

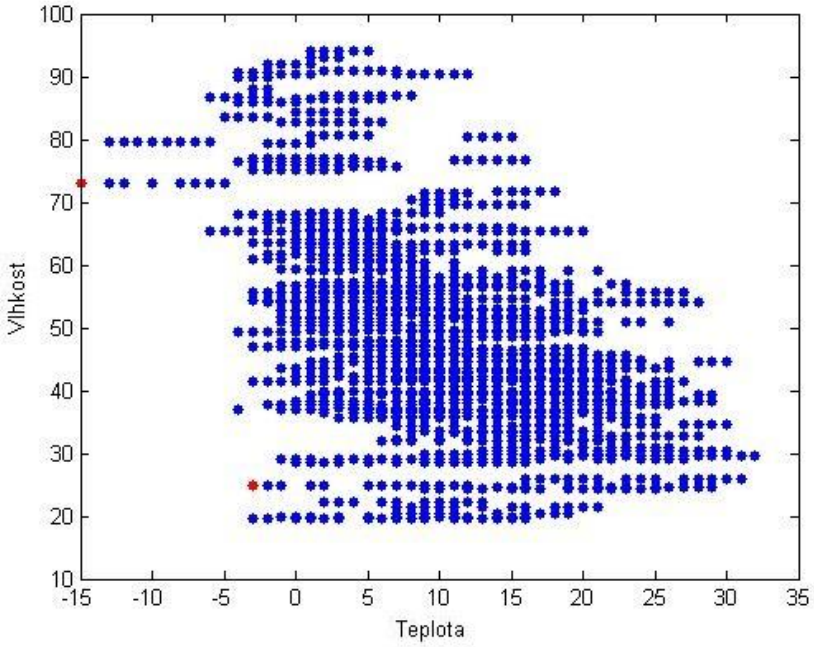
Také je možné rozlišit jednotlivé typy poruch a názorně je vyobrazit v *Obrázku 4.6.*



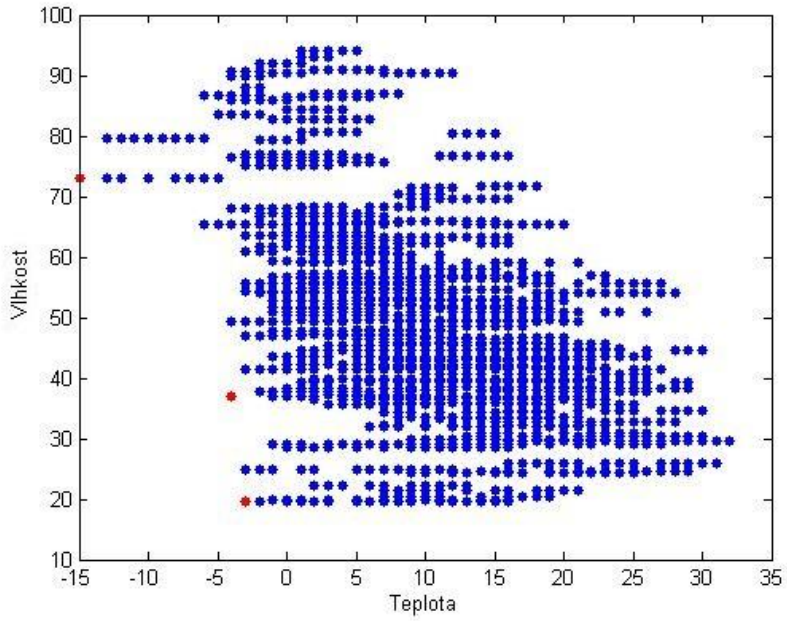
Obrázek 4.6: Zobrazení vícerozměrných reálných dat s typy poruch

Z *Obrázku 4.6* opět nelze usuzovat, že by se data řídila nějakým známým vícerozměrným rozdělením, a tak je možné aplikovat pouze neparametrické metody detekce odlehlých hodnot ve vícerozměrných souborech dat.

První takovouto metodou je metoda *KDIST*, případně *MeanDIST*, které jsou založené na vzdálenostech *k*-nejbližších sousedů. Tyto metody budou aplikovány nejdříve na data, ve kterých nebyly nalezeny žádné poruchy a následně na poruchová měření. Toto rozdělení dat pak bude využito i při ostatních metodách detekce outliers. Metodami *KDIST* a *MeanDIST*, které hledaly *5*-nejbližší sousedy každého bodu a jsou založeny na výpočtu Euklidovských vzdáleností, jsou získány výsledky zobrazené v následujících *Obrázcích 4.7* a *4.8*. Zatímco metodou *KDIST* jsou nalezeny vždy stejné odlehlé hodnoty při jakékoliv volbě parametru t, metodou *MeanDIST* jsou nalezeny při volbách $t = 0{,}5$ a $0{,}6$ čtyři outliers a při volbě například $t = 0{,}7$ pak dva outliers.



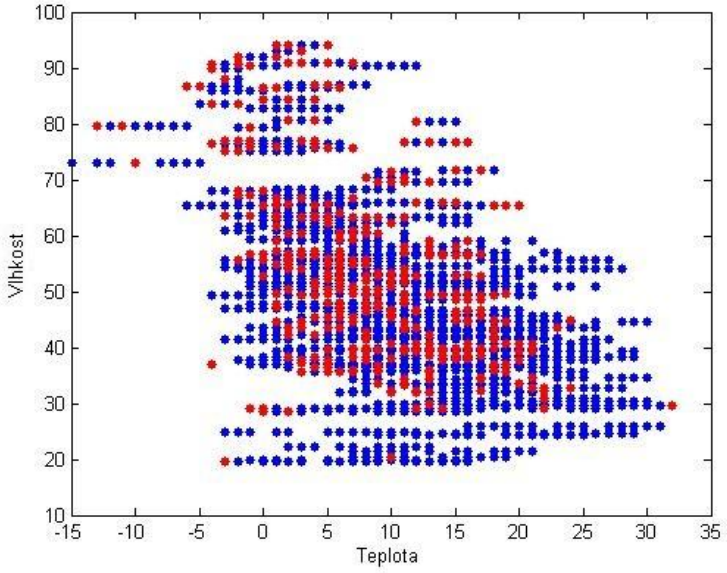
Obrázek 4.7: Detekované outliers metodou KDIST



Obrázek 4.8: Detekované outliers metodou MeanDIST, $t = 0,6$

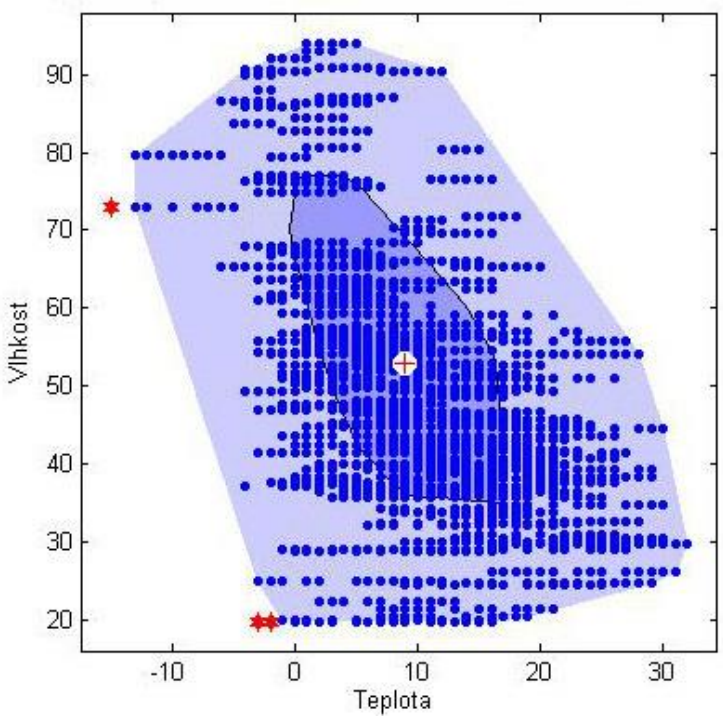
Výsledky získané metodami KDIST a MeanDIST lze srovnat také s výsledky získanými pomocí jiných postupů. Další metodou (neparametrickou), jak detekovat odlehlé hodnoty vícerozměrných dat, je tzv. LOF metoda, která je založená na porovnávání hustot bodů a jejich *k*-nejbližších sousedů. Tato metoda však není příliš vhodná pro detekci outliers v tomto souboru dat, protože je zde mnoho totožných hodnot, což způsobí, že nulové vzdálenosti *k*-nejbližšího souseda dále ovlivňují výpočet hodnot lrd jednotlivých bodů. A to tím způsobem, že hodnoty lrd nabývají nekonečna, což ovšem

také vede k nekonečnu v potřebném ukazateli LOF, pomocí které lze detekovat outliers. Takto však tyto body detekovat nelze, protože pokud by opět platila mez identifikace outliers vyšší než 1, při volbě meze 1,2 by metoda detekovala hned 1073 odlehlých hodnot, které jsou zobrazeny na *Obrázku 4.9.* Většina z nich je však detekována chybně z důvodu duplicity bodů.



Obrázek 4.9: Detekované „outliers" metodou LOF

Jako poslední je možné na data aplikovat bagplot, který také znázorní odlehlé body.

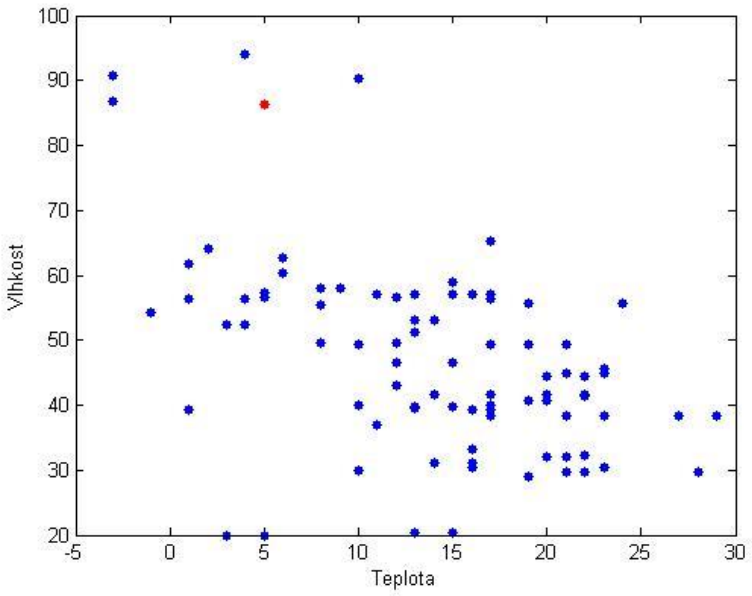


Obrázek 4.10: Bagplot dat bez poruch

Porovnáním výsledků metodou bagplot spolu s metodami *KDIST* a *MeanDIST* jsou jako odlehlé hodnoty detekovány takové kontrolní body, kdy byla naměřena nízká teplota a vlhkost pod 20 % nebo naopak přes 70 %.

Nyní je možné detekovat ještě outliers v datech, kdy nastala porucha. Těchto hodnot bylo v datech nalezeno celkem 87. Opět budou aplikovány metody *KDIST, MeanDIST* a na závěr bagplot. Jelikož je zde menší počet zkoumaných hodnot, bude uplatněna také metod LOF, která v předešlém případě selhala kvůli velkému množství duplicitních hodnot.

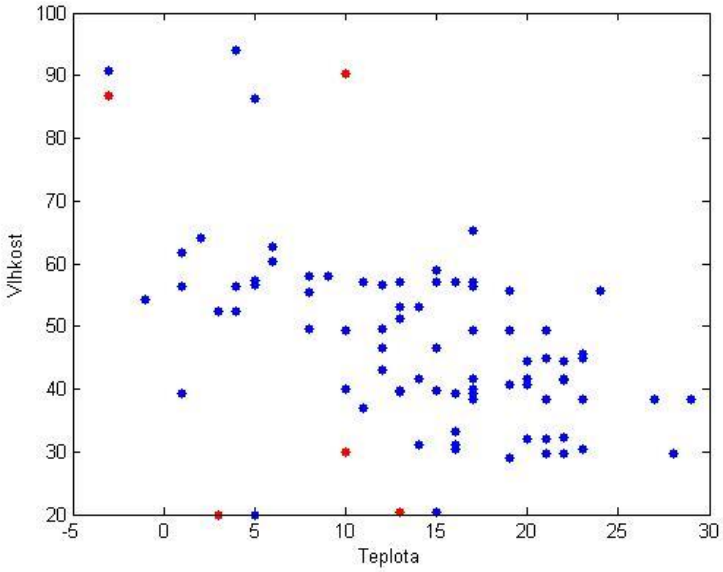
Ovšem reálná data strojů, u kterých byla objevena porucha, nejsou příliš kompaktní a nelze také s přesností tvrdit, že neobsahují *Masking effect* nebo *Swamping effect*. Právě z tohoto důvodu je možné, že metody nebudou fungovat přesně, jak je potřeba. Navíc u metod *KDIST* a *MeanDIST* nastává problém, který byl zmíněný již při numerických experimentech, a to podobná hodnota *KDIST* (případně *MeanDIST*), což zapříčiňuje nepatrnou diferenci mezi těmito body a jako outlier bude detekován pouze jeden z nich. S touto znalostí je možné zobrazit detekované odlehlé hodnoty metodou *KDIST* na *Obrázku 4.11*, kdy počet outliers se opět nemění v závislosti na voleném parametru t.



Obrázek 4.11: Detekovaný outlier dat poruch metodou KDIST

Na *Obrázku 4.11* je patrné, že se zde objevuje menší skupina dat v levém horním rohu grafu, která může ovlivňovat celý soubor a následně tak také detekci odlehlých hodnot.

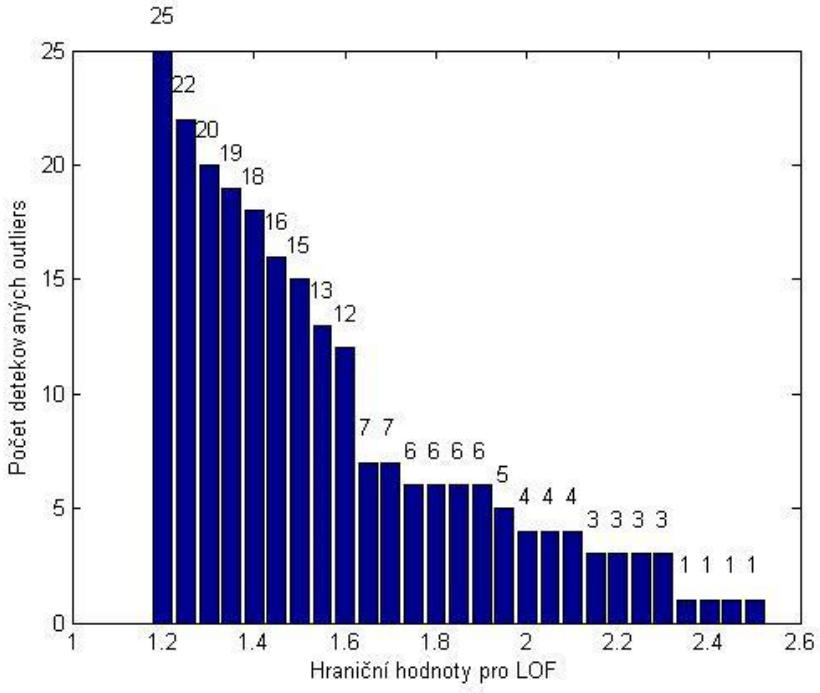
Metodou *MeanDIST* při volbě parametru $t = 0{,}5$ pak byly detekovány jako odlehlé hodnoty body zobrazené na následujícím *Obrázku 4.12*.



Obrázek 4.12: Detekované outliers dat poruch metodou MeanDIST, $t = 0,5$

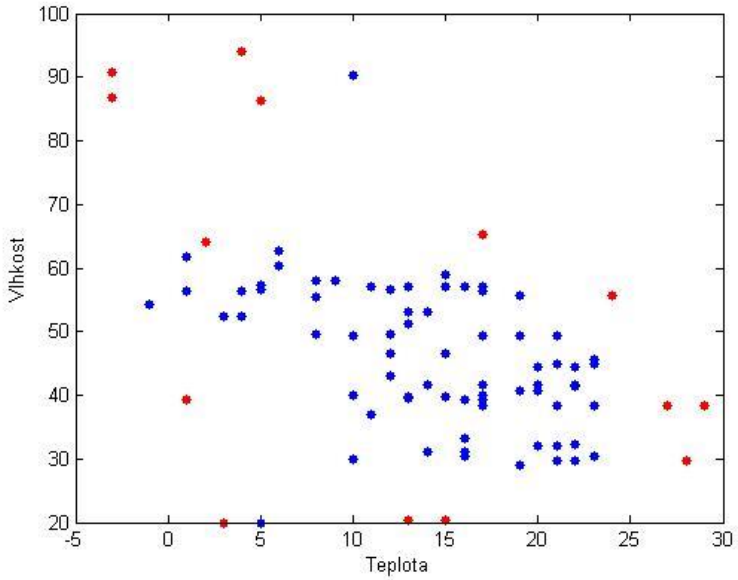
S rostoucím parametrem t pak ubývají identifikované odlehlé hodnoty, a to tak, že jako outliers zůstávají detekované hodnoty, jejichž vlhkost odpovídá 20 %.

Zde již může být uplatněna také metoda LOF, neboť nedochází k žádným duplicitním bodům, a tím pádem k výpočtům nekonečna ve výsledném místním faktoru odlehlosti. Opět však záleží na volbě, kdy je pozorování považováno za odlehlé. Počty detekovaných outliers v závislosti na volbě hranice je uvedena v následujícím *Obrázku 4.13.*

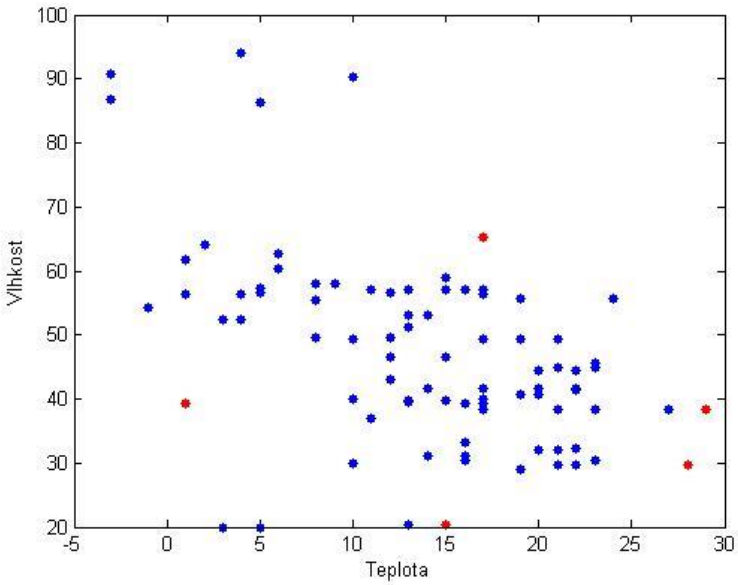


Obrázek 4.13: Počty detekovaných outliers metodou LOF v závislosti na volbě hranice

Při volbě hranice 1,5 jsou tak jako odlehlé hodnoty určeny body uvedené na *Obrázku 4.14*, zatímco při volbě hranice 1,8 je jich o poznání méně a jsou uvedené na *Obrázku 4.15.*



Obrázek 4.14: Detekované outliers metodou LOF při hranici 1,5



Obrázek 4.15: Detekované outliers metodou LOF při hranici 1,8

Pokud jde tedy o tato data, bylo by vhodné překontrolovat 5 hodnot tvořících menší shluk v levém horním rohu grafu, aby nebyly metody zkreslovány případnými efekty. Případně by bylo možné všechny tyto hodnoty považovat za odlehlé. Vše ovšem záleží na pozorovateli a způsobu, jak vyhodnotí, zda zvolené hodnoty vyřadí z dalších analýz nebo je ponechá jako extrémní hodnoty, které nesou důležitou informaci o souboru dat.

5 Závěr

Cílem této diplomové práce bylo prozkoumat a popsat různé metody identifikace odlehlých pozorování, které se využívají v nejrůznějších případech, kdy je potřebné odhalit možné chybné hodnoty v datových souborech. V současné době dochází k analyzování velkého množství dat. Z těchto dat pak mohou být tvořeny predikce budoucího vývoje, a tak je velmi důležité, aby byla data správná, bez jakýchkoliv chyb způsobených lidským faktorem nebo například poruchou přístrojů využitých ke sběru a zpracování dat. Detekce outliers je tak zásadní disciplínou předcházející dalšímu statistickému zpracování získaných dat.

V práci byly popsány nejrůznější metody detekce odlehlých hodnot. Nejdříve metody, které je možné využívat pro identifikaci outliers v jednorozměrných datech, jako jsou Grubbsův test, Dean-Dixonův test a grafická metoda – boxploty. Grubbsův a Dean-Dixonův test jsou parametrickými metodami předpokládající normální rozdělení souboru dat, přičemž volba vhodného testu závisí také na množství testovaných hodnot.

Následovaly metody vhodné pro vícerozměrná data. Ty mohou být rozděleny na distance-based metody a density-based metody. Detailně popsány a následně využity byly metody *KDIST* a *MeanDIST*, jako zástupci metod *k*-nejbližších sousedů, a následně také *Local outlier factor (LOF)*, jako zástupce metod založených na porovnání hustot bodu a jeho *k*-nejbližších sousedů. Doplněny byly grafickou metodou – bagplot a detekcí odlehlých bodů pomocí Mahalanobisovy vzdálenosti.

V úvahu byly brány také časové řady a detekce outliers v regresi. Jelikož má mnoho zkoumaných datových souborů časový charakter, není vhodné opomenout ani tuto možnost. Pro detekci tzv. vlivných bodů se využívá více metod. V regresi jsou navíc rozlišovány odlehlé hodnoty ve směru osy x (leverage points) a ve směru osy y (outliers). Pro identifikaci leverage points byla využita metoda založená na porovnávání diagonálních prvků projekční matice H. Pro identifikaci outliers pak Cookova vzdálenost a Welsch-Kuhova vzdálenost. Tyto metody byly doplněny opět grafickou metodou – Williamsovým grafem.

Pro lepší názornost a princip fungování jednotlivých metod byly provedeny numerické experimenty v programu Matlab. Vždy byly vygenerovány datové soubory dle potřebných předpokladů pro jednotlivé metody detekce outliers a následně zde byly záměrně poškozeny některé z hodnot nebo byly do souboru záměrně odlehlé hodnoty přidány. Tyto numerické experimenty také odhalily výhody a nevýhody jednotlivých metod. Pokud jde například o metody jednorozměrných dat, jako jsou Grubbsův a Dean-Dixonův test, jako nevýhoda se jeví předpoklad

normálního rozdělení. Dojde-li totiž k přidání záměrně odlehlých hodnot k souboru dat o deseti pozorováních, může se stát, že normalita nezůstane zachována. Z tohoto důvodu je vhodné doplnit metodu ještě neparametrickými boxploty.

Při detekci odlehlých pozorování ve vícerozměrných datech metodami *KDIST* a *MeanDIST* je největší nevýhodou, existuje-li v souboru více potenciálně odlehlých hodnot, kterým je však přiřazena podobná hodnota *KDIST* (případně *MeanDIST*). V tomto případě je totiž diference seřazených hodnot minimální a body nemusí být odhalené jako odlehlé. Postup detekce outliers je tak vhodné opakovat. Další nevýhodou je volba parametru $t \in (0,1)$, jehož hodnota ovlivňuje počet detekovaných outliers a je zcela na pozorovateli.

Metoda LOF detekující odlehlé hodnoty na základě porovnávání hustot bodu a jeho *k*-nejbližších sousedů je pak velmi citlivá na výskyt *Masking effectu* nebo *Swamping effectu.* Hlavní nevýhoda této metody však nastává, pokud se v souboru dat vyskytnou totožná pozorování, což nastalo také v případě reálných dat. Metody detekce outliers vícerozměrných dat byly také doplněny grafickou metodou – bagplot.

Na závěr práce byly zkoumané metody aplikovány na reálná data, kde bylo zjištěno, že odlehlé hodnoty se nachází v datech v případě poruch strojů. Konkrétně pak v záznamech, při jakých hodnotách vlhkosti dochází k poruchám strojů, kde byly detekované outliers při vlhkosti kolem 90 %. Budou-li data chápána jako vícerozměrná, pak lze identifikovat outliers také při závislosti vlhkosti na teplotě a na pozorovateli tak záleží, zda některé detekované hodnoty vyhodnotí jako extrémy nebo jako odlehlé hodnoty, a bude dále řešit, jak se s nimi vypořádat.

Použitá literatura a zdroje

[1] BARNETT, Vic a Toby LEWIS. *Outliers in statistical data*. New York: Wiley, 1978, ISBN 0471995991.

[2] Box plot. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001-2015 [cit. 2015-02-10]. Dostupné z: http://en.wikipedia.org/wiki/Box_plot

[3] Normální rozdělení: Gaussova křivka. In: *WikiSkripta* [online]. 2013 [cit. 2015-02-10]. Dostupné z: http://www.wikiskripta.eu/index.php/Norm%C3%A1ln%C3%AD_rozd%C4%9Blen%C3%AD

[4] Analyzing data: Box and whisker plot. In: *MathCaptain* [online]. 2014 [cit. 2015-02-18]. Dostupné z: http://www.mathcaptain.com/statistics/analyzing-data.html

[5] Grubbs' outlier test. In: *Fundamentals of statistics* [online]. 2012 [cit. 2015-02-18]. Dostupné z: http://www.statistics4u.com/fundstat_eng/ee_grubbs_outliertest.html

[6] Outlier test - Dean and Dixon. In: *Fundamentals of statistics* [online]. 2012 [cit. 2015-02-18]. Dostupné z: http://www.statistics4u.com/fundstat_eng/cc_outlier_tests_dixon.html

[7] MELOUN, Milan a Jiří MILITKÝ. *Interaktivní statistická analýza dat*. Vyd. 3., V nakl. Karolinum 1. Praha: Karolinum, 2012. ISBN 978-80-246-2173-9.

[8] ORAIR, Gustavo H., Carlos H. C. TEIXEIRA, Wagner MEIRA JR., Ye WANG a Srinivasan PARTHASARATHY. *Distance-Based Outlier Detection: Consolidation and Renewed Bearing.* The Ohio State University, Columbus, USA [online]. [cit. 2015-02-22]. Dostupné z: http://www.vldb.org/pvldb/vldb2010/papers/I09.pdf

[9] BAY, Stephen D. a Mark SCHWABACHER. *Near Linear Time Detection of Distance-Based Outliers and Applications to Security*. California, USA [online]. [cit. 2015-02-22]. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.119.6186&rep=rep1&type=pdf

[10] HAUTAMÄKI, Ville, Ismo KÄRKKÄINEN a Pasi FRÄNTI. *Outlier Detection Using k-Nearest Neighbour Graph.* Joensuu, Finland [online]. [cit. 2015-02-22]. Dostupné z: ftp://ftp.cs.joensuu.fi/pub/franti/papers/Hautamaki/P2.pdf

[11] Local outlier factor. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001-2015 [cit. 2015-03-08]. Dostupné z: http://en.wikipedia.org/wiki/Local_outlier_factor

[12] BREUNIG, Markus M., Hans-Peter KRIEGEL, Raymond T. NG a Jörg SANDER. *LOF: Identifying Density-Based Local Outliers.* Munich, Germany; Vancouver, Canada [online]. [cit. 2015-03-10]. Dostupné z: http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf

[13] GraphPad Software. *Detecting outliers with Grubbs' test* [online]. 2010 [cit. 2015-03-10]. Dostupné z: http://graphpad.com/support/faqid/1598/

[14] MELOUN, Milan a Jiří MILITKÝ. *Kompendium statistického zpracování dat: metody a řešené úlohy včetně CD*. Vyd. 1. Praha: Academia, 2002. ISBN 80-200-1008-4.

[15] BEN-GAL, Irad. *OUTLIER DETECTION: Chapter 1* [online]. Tel-Aviv University: Kluwer Academic Publishers, Israel, 2005 [cit. 2015-04-10]. ISBN 0387244352. Dostupné z: http://www.eng.tau.ac.il/~bengal/outlier.pdf

[16] DEVORE, Jay L a Roxy PECK. *Statistics: the exploration and analysis of data*. St. Paul: West Pub. Co., 1986. ISBN 0314931724.

[17] Delete-1 Statistics: DFFITS. In: *MathWorks* [online]. 1994-2015 [cit. 2015-04-25]. Dostupné z: http://www.mathworks.com/help/stats/delete-1-statistics.html

[18] ROUSSEEUW, Peter J., Ida RUTS a John TUKEY. The Bagplot: A Bivariate Boxplot. *The American Statistician* [online]. 1999, s. 382-387 [cit. 2015-05-02]. Dostupné z: http://venus.unive.it/romanaz/ada2/bagplot.pdf

[19] BLATNÁ, Dagmar. *OUTLIERS IN REGRESSION* [online]. University of Economics Prague [cit. 2015-03-10]. Dostupné z: http://statistika.vse.cz/konference/amse/PDF/Blatna.pdf

[20] CHATTERJEE, Samprit, Ali S HADI a Bertram PRICE. *Regression analysis by example*. 4. vydání. New York: Wiley, 2006, str. 90. Wiley series in probability and statistics. ISBN 0471319465.

[21] StatisticalHelp. MULTIPLE (GENERAL) LINEAR REGRESSION. [online]. 2000-2015 [cit. 2015-03-10]. Dostupné z: http://www.statsdirect.com/help/default.htm#regression_and_correlation/multiple_linear.htm

[22] FORBELSKÁ, Marie. Math.muni. METODY REGRESNÍ DIAGNOSTIKY. [online]. 2000-2015 [cit. 2015-03-10]. Dostupné z: http://www.math.muni.cz/~kolacek/docs/frvs/M6120/materialy/M6120cv08.pdf

[23] TriloByte. LINEÁRNÍ REGRESE. [online]. 2015 [cit. 2015-03-10]. Dostupné z: http://www.trilobyte.cz/downloadfree/qcemanual/linreg.pdf

[24] SUÁREZ RANCEL, M. Mercedes a Miguel A. GONZÁLES SIERRA. *Weighting and Deletion Approaches to Regression Diagnostics: A Comparison and an Extension.* [online]. 2000 [cit. 2015-04-10]. Dostupné z: http://jesr.journal.fatih.edu.tr/weighting-mercedes.pdf

[25] CHATTERJEE, Samprit, Ali S HADI a Bertram PRICE. *Sensitivity Analysis in Linear Regression*. New York: Wiley, 2006. Dostupné z:

https://books.google.cz/books?id=qozaTQMdC0EC&pg=PA120&lpg=PA120&dq=welsch+kuh+distance&source=bl&ots=nn2O9rml3x&sig=BxHp--Ij4nfkCeSDhYQGYzc5yaE&hl=cs&sa=X&ei=1Cb8VOH6PMvcPbKcgZgB&ved=0CCYQ6AEwAQ#v=onepage&q=welsch%20kuh%20distance&f=false

[26] ANTOCH, Jaromír a Dana VORLÍČKOVÁ. *Vybrané metody statistické analýzy dat*. 1. vyd. Praha: Academia, 1992. ISBN 8020002049.

[27] CHATTERJEE, Samprit a Ali S HADI. *Sensitivity analysis in linear regression*. New York: Wiley, 1988. ISBN 0471822167.

[28] KU LEUVEN. *LIBRA files: LIBRA31okt11* [online]. 2008 [cit. 2015-05-03]. Dostupné z: https://wis.kuleuven.be/stat/robust/Programs/LIBRA

[29] kNN graph. In: *Google.com* [online]. Původní obrázek ve formátu PNG [cit. 2015-04-15]. Dostupné z: http://3.bp.blogspot.com/-QGJlQDEHo_0/UY2VytMaQ_I/AAAAAAAAAlk/sHg5pNc3aP4/s1600/SGPlot4.png

Přílohy

Příloha 1. – Tabulka kritických hodnot T_α pro Grubbsův test

	Kritické hodnoty pro Grubbsův test	
N \ α	5,00%	1,00%
3	1,153	1,155
4	1,463	1,493
5	1,671	1,749
6	1,822	1,944
7	1,938	2,097
8	2,032	2,221
9	2,110	2,323
10	2,176	2,410
11	2,234	2,484
12	2,285	2,549
13	2,331	2,607
14	2,372	2,658
15	2,409	2,705
16	2,443	2,747
17	2,475	2,785
18	2,504	2,821
19	2,531	2,853
20	2,557	2,884
30	2,745	3,103
40	2,868	3,239
50	2,957	3,337
60	3,027	3,411
70	3,084	3,471
80	3,132	3,521
90	3,173	3,563
100	3,210	3,600

Tabulka Přílohy.1: Tabulka kritických hodnot pro Grubbsův test (Zdroj [5], vlastní zpracování)

Příloha 2. – Tabulka kritických hodnot Q_α pro Dean-Dixonův test

N \ α	1,00%	2,00%	5,00%	10,00%	20,00%
3	0,988	0,976	0,941	0,886	0,782
4	0,889	0,847	0,766	0,679	0,561
5	0,782	0,729	0,643	0,559	0,452
6	0,698	0,646	0,563	0,484	0,387
7	0,636	0,587	0,507	0,433	0,344
8	0,591	0,542	0,467	0,398	0,314
9	0,555	0,508	0,436	0,370	0,291
10	0,527	0,482	0,412	0,349	0,274

Tabulka Přílohy.2: Tabulka kritických hodnot pro Dean-Dixonův test (Zdroj [6], vlastní zpracování)

Příloha 3. – Dosažitelné vzdálenosti zvolených bodů

B\A	1	2	3	4	5	6	7	8	9	10
1	-	1,342	1,342	1,342	1,697	1,720	1,924	2,283	2,280	3,002
2	1,118	-	0,640	0,640	0,728	0,640	0,806	1,166	1,204	1,887
3	1,140	0,583	-	0,583	0,583	0,583	0,825	1,170	1,140	1,879
4	1,342	0,447	0,447	-	0,600	0,447	0,583	0,943	1,000	1,664
5	1,697	0,728	0,583	0,600	-	0,583	0,583	0,806	0,632	1,432
6	1,720	0,640	0,583	0,447	0,447	-	0,447	0,608	0,566	1,300
7	1,924	0,806	0,825	0,583	0,583	0,510	-	0,510	0,510	1,082
8	2,283	1,166	1,170	0,943	0,806	0,608	0,608	-	0,608	0,721
9	2,280	1,204	1,140	1,000	0,632	0,566	0,566	0,566	-	0,806
10	3,002	1,887	1,879	1,664	1,432	1,300	1,082	1,082	1,082	-

Tabulka Přílohy.3: Tabulka dosažitelných vzdáleností zvolených bodů metody LOF

Příloha 4. – Místní faktory odlehlosti (LOF) všech pozorování

Pozorování	1	2	3	4	5	6	7	8	9	10	11	12	13	14
LOF	1,085	1,276	1,205	1,020	0,993	1,027	1,103	0,996	2,399	1,184	1,078	1,175	0,954	1,043
Pozorování	15	16	17	18	19	20	21	22	23	24	25	26	27	28
LOF	1,917	1,024	0,987	1,044	1,062	0,990	1,545	0,927	1,016	1,237	1,052	1,000	1,499	1,015
Pozorování	29	30	31	32	33	34	35	36	37	38	39	40	41	42
LOF	1,391	1,459	0,983	0,981	0,968	1,126	2,247	0,966	1,322	1,451	1,503	0,998	1,126	1,008
Pozorování	43	44	45	46	47	48	49	50	51	52	53	54	55	56
LOF	1,016	1,048	1,033	1,236	0,956	1,042	1,034	1,068	1,066	1,019	1,019	1,581	1,120	0,990
Pozorování	57	58	59	60	61	62	63	64	65	66	67	68	69	70
LOF	1,016	0,947	1,011	1,039	1,071	0,999	1,176	1,033	0,974	1,019	0,986	1,036	0,990	0,982
Pozorování	71	72	73	74	75	76	77	78	79	80	81	82	83	84
LOF	1,067	0,990	1,025	1,074	0,943	1,106	1,003	0,983	1,021	0,976	1,324	1,043	0,996	1,061
Pozorování	85	86	87	88	89	90	91	92	93	94	95	96	97	98
LOF	1,086	1,080	1,067	1,044	1,177	1,070	0,937	1,024	1,039	1,812	1,022	1,066	1,066	1,108
Pozorování	99	100	101	102	103	104	105	106	107	108	109	110	111	112
LOF	0,996	1,008	1,030	1,083	1,071	1,030	1,020	1,008	1,198	1,059	1,010	1,013	1,124	1,034
Pozorování	113	114	115	116	117	118	119	120	121	122	123	124	125	126
LOF	1,076	1,007	0,971	1,238	1,026	0,962	1,024	0,966	0,996	1,053	1,145	0,931	1,061	1,034
Pozorování	127	128	129	130	131	132	133	134	135	136	137	138	139	140
LOF	1,138	1,001	1,141	0,994	1,029	1,011	1,063	1,065	1,244	0,958	1,396	0,970	1,003	0,935
Pozorování	141	142	143	144	145	146	147	148	149	150	151	152	153	154
LOF	1,022	0,975	1,003	0,964	0,979	1,424	1,015	1,436	0,981	1,003	1,167	1,015	1,111	1,079

Tabulka Přílohy.4: Místní faktory odlehlosti - 1.část

Pozorování	155	156	157	158	159	160	161	162	163	164	165	166	167	168
LOF	1,084	1,001	0,991	1,058	0,961	1,391	1,095	1,524	0,980	1,039	1,992	1,204	0,983	1,026
Pozorování	169	170	171	172	173	174	175	176	177	178	179	180	181	182
LOF	0,960	0,966	1,018	1,315	1,015	0,991	0,957	0,990	0,950	1,273	1,153	0,998	1,066	0,990
Pozorování	183	184	185	186	187	188	189	190	191	192	193	194	195	196
LOF	0,973	0,972	1,157	1,086	0,996	1,035	1,017	1,014	1,009	0,954	0,975	1,083	1,120	0,993
Pozorování	197	198	199	200	201	202	203	204	205	206	207	208	209	210
LOF	1,022	1,737	1,033	1,183	0,982	1,009	0,973	1,022	1,011	0,975	1,007	1,832	1,000	1,021
Pozorování	211	212	213	214	215	216	217	218	219	220	221	222	223	224
LOF	1,017	1,099	0,969	1,079	1,070	1,050	0,990	1,220	1,133	1,055	0,987	1,959	1,021	1,062
Pozorování	225	226	227	228	229	230	231	232	233	234	235	236	237	238
LOF	1,165	1,008	1,320	1,003	0,962	1,121	1,069	0,993	1,005	1,060	1,103	1,070	1,116	1,569
Pozorování	239	240	241	242	243	244	245	246	247	248	249	250	251	252
LOF	1,150	1,029	1,067	1,004	1,140	1,056	0,969	1,061	1,047	0,989	1,257	0,994	0,995	1,109
Pozorování	253	254	255	256	257	258	259	260	261	262	263	264	265	266
LOF	0,969	1,016	1,033	0,962	0,933	1,018	0,946	0,993	1,030	1,508	1,663	1,013	1,005	1,044
Pozorování	267	268	269	270	271	272	273	274	275	276	277	278	279	280
LOF	1,010	1,062	1,321	1,462	1,055	1,004	0,946	1,420	0,978	0,987	1,711	1,053	0,935	1,041
Pozorování	281	282	283	284	285	286	287	288	289	290	291	292	293	294
LOF	1,043	0,978	0,969	1,773	0,980	1,163	1,123	1,056	1,174	1,099	1,159	0,991	1,115	1,001
Pozorování	295	296	297	298	299	300								
LOF	0,982	0,996	1,093	1,176	1,249	1,148								

Tabulka Přílohy.5: Místní faktory odlehlosti - 2.část