

Statistická analýza složek kvality bílého vína

Petr Voborník

Fakulta informatiky a managementu, Katedra informatiky a kvantitativních metod
Univerzita Hradec Králové, Rokitanského 62, 500 03 Hradec Králové, Česká republika
petr.vobornik@uhk.cz

Úvod

Data set `winequality-white.csv` byl stažen z webu UCI Machine Learning Repository ⁽¹⁾. Obsahuje 4898 záznamů o měření kvality bílé varianty portugalského vína „Vinho Verde“¹, přičemž pokaždé bylo zaznamenáno 12 hodnot týchž parametrů, majících vliv na celkovou kvalitu vína. Veškeré hodnoty vždy byly objektivně změřeny, kvalitu pak tvoří celé číslo od 0 (velmi špatné) do 10 (vynikající) určené jako medián z nezávislého hodnocení minimálně tří vinařských odborníků.



Index	Proměnná	Název	Popis
1	fixed_acid	fixed acidity	obsah netěkavých kyselin
2	volatile_acid	volatile acidity	obsah těkavých kyselin
3	citric_acid	citric acid	kyselina citrónová
4	resid_sugar	residual sugar	zbytkový cukr
5	chlorides	chlorides	chloridy
6	free_sdo	free sulfur dioxide	volný oxid siřičitý
7	total_sdo	total sulfur dioxide	celkový obsah oxidu siřičitého
8	density	density	hustota
9	pH	pH	pH
10	sulphates	sulphates	sírany
11	alcohol	alcohol	alkohol
12	quality	quality	kvalita

Tabulka 1 – Popis proměnných.

Bohužel, kvůli zachování obchodního tajemství, jsou k dispozici pouze fyzikálně-chemické (vstupy) a senzorické (výstupy) proměnné a údaje o např. typech hroznů, značce vína, prodejní ceně apod. chybí. V testech byla zastoupena především normální vína, kvalitní a nekvalitní se vyskytují jen zřídka. Ne všechny proměnné zároveň musí být pro kvalitu vína zcela relevantní. ⁽¹⁾

Pokusíme se nad těmito daty provést vícerozměrnou regresní analýzu a najít jaký vliv mají hodnoty jednotlivých proměnných (složek vína) na jeho celkovou kvalitu.

Použitý software

Na prvotní úpravu dat, základní a pomocné výpočty byl použit [Microsoft Excel 2007](#). Pro statistické výpočty byla použita trial verze statistického software [Statistica 9²](#). Veškeré zde uvedené grafy, výstupy a tabulky jsou, není-li uvedeno jinak, převzaty právě z programu Statistica.

¹ <http://www.vinhoverde.pt/en/>

² <http://www.statsoft.cz/podpora/ke-stazeni/trial-verze-statistica/>

Statistiky

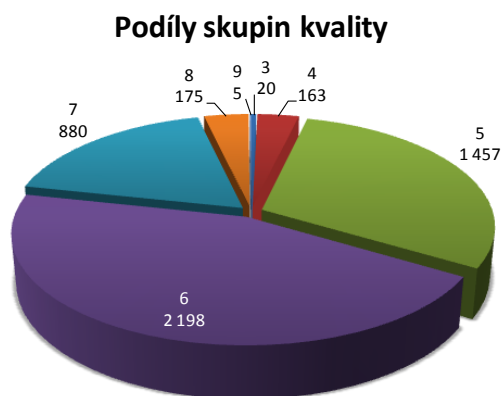
Popisné statistiky

Veškeré datové údaje všech proměnných jsou číselné. Proměnné `free_sdo` (volný oxid siřičitý), `total_sdo` (celkový obsah oxidu siřičitého) a `quality` (kvalita) jsou celočíselné.

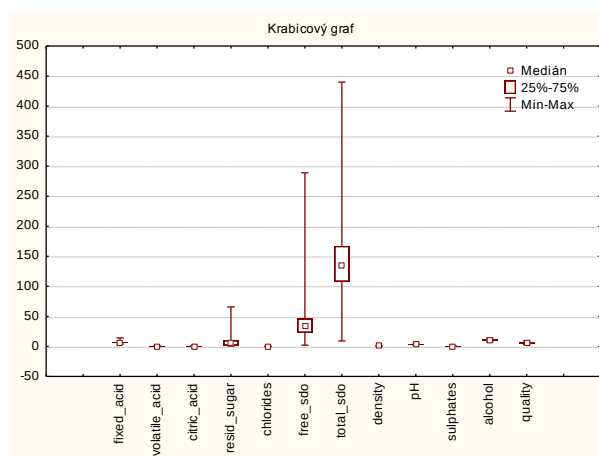
	Průměr	Medián	Minimum	Maximum	Rozptyl	Sm.odch.	Var.koef.	Šikmost	Špičatost
<code>fixed_acid</code>	6,855	6,800	3,800	14,200	0,712	0,8439	12,3106	0,6478	2,1722
<code>volatile_acid</code>	0,278	0,260	0,080	1,100	0,010	0,1008	36,2256	1,5770	5,0916
<code>citric_acid</code>	0,334	0,320	0,000	1,660	0,015	0,1210	36,2127	1,2819	6,1749
<code>resid_sugar</code>	6,391	5,200	0,600	65,800	25,726	5,0721	79,3574	1,0771	3,4698
<code>chlorides</code>	0,046	0,043	0,009	0,346	0,000	0,0218	47,7318	5,0233	37,5646
<code>free_sdo</code>	35,308	34,000	2,000	289,000	289,243	17,0071	48,1678	1,4067	11,4663
<code>total_sdo</code>	138,361	134,000	9,000	440,000	1 806,085	42,4981	30,7154	0,3907	0,5719
<code>density</code>	0,994	0,994	0,987	1,039	0,000	0,0030	0,3009	0,9778	9,7938
<code>pH</code>	3,188	3,180	2,720	3,820	0,023	0,1510	4,7361	0,4578	0,5308
<code>sulphates</code>	0,490	0,470	0,220	1,080	0,013	0,1141	23,2983	0,9772	1,5909
<code>alcohol</code>	10,514	10,400	8,000	14,200	1,514	1,2306	11,7043	0,4873	-0,6984
<code>quality</code>	5,878	6,000	3,000	9,000	0,784	0,8856	15,0672	0,1558	0,2165

Tabulka 2 – Popisné statistiky proměnných.

Základní popisné statistiky lépe charakterizují vstupní data jednotlivých proměnných (Tabulka 2). Je z nich patrný jejich rozsah, průměrná hodnota, rozptyl apod. Díky koeficientům šikmosti a špičatosti si lze udělat představu i o tvaru jejich normálního průběhu.



Graf 1 – Výšečový graf podílu počtu jednotlivých skupin kvality v testovaných vzorcích (Excel).



Graf 2 – Krabicový graf hodnot jednotlivých proměnných.

Výšečový Graf 1 zobrazuje četnosti zastoupení jednotlivých skupin kvality v datovém souboru. Například kvalitativní skupiny 9 bylo dosaženo pouze v pěti případech, naopak nejčetnější zastoupení má skupina 6, kteréžto hodnocení bylo vínu přiděleno v 2 198 případech.

Krabicový Graf 2 pro přehlednost graficky znázorňuje rozsahy hodnot jednotlivých proměnných.

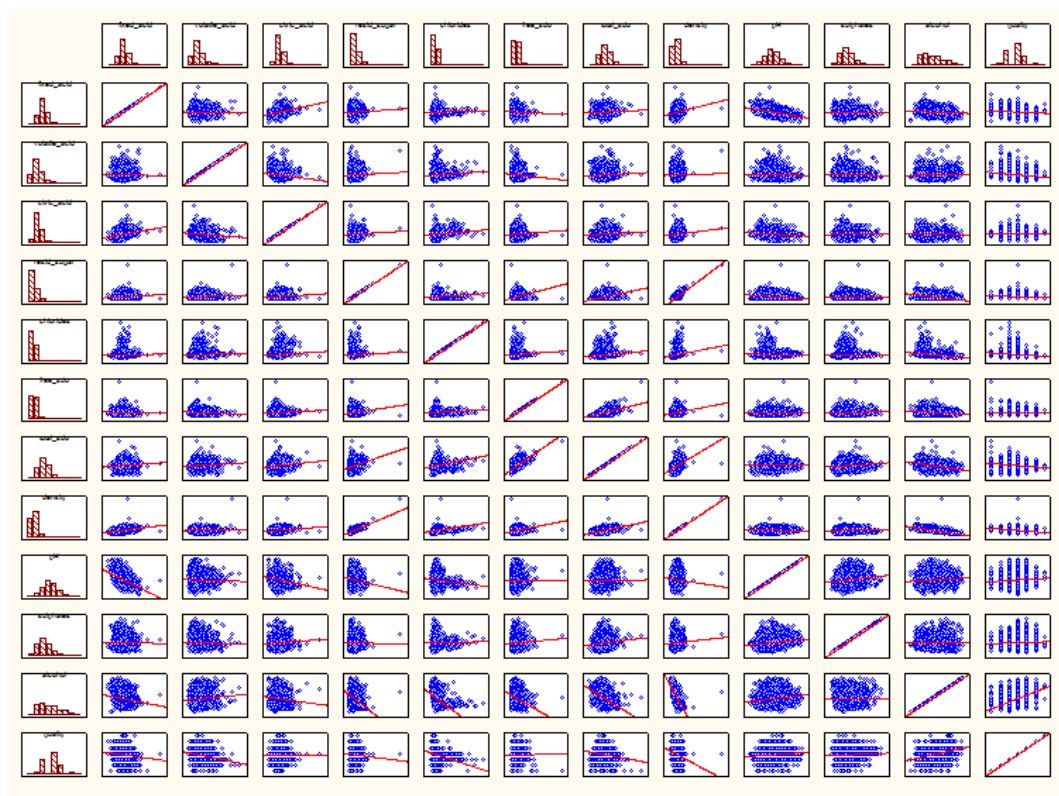
Korelační matice a multikolinearita

	fixed_acid	volatile_acid	citric_acid	resid_sugar	chlorides	free_sdo	total_sdo	density	pH	sulphates	alcohol	quality
fixed_acid	1,000	-0,023	0,289	0,089	0,023	-0,049	0,091	0,265	-0,426	-0,017	-0,121	-0,114
volatile_acid	-0,023	1,000	-0,149	0,064	0,071	-0,097	0,089	0,027	-0,032	-0,036	0,068	-0,195
citric_acid	0,289	-0,149	1,000	0,094	0,114	0,094	0,121	0,150	-0,164	0,062	-0,076	-0,009
resid_sugar	0,089	0,064	0,094	1,000	0,089	0,299	0,401	0,839	-0,194	-0,027	-0,451	-0,098
chlorides	0,023	0,071	0,114	0,089	1,000	0,101	0,199	0,257	-0,090	0,017	-0,360	-0,210
free_sdo	-0,049	-0,097	0,094	0,299	0,101	1,000	0,616	0,294	-0,001	0,059	-0,250	0,008
total_sdo	0,091	0,089	0,121	0,401	0,199	0,616	1,000	0,530	0,002	0,135	-0,449	-0,175
density	0,265	0,027	0,150	0,839	0,257	0,294	0,530	1,000	-0,094	0,074	-0,780	-0,307
pH	-0,426	-0,032	-0,164	-0,194	-0,090	-0,001	0,002	-0,094	1,000	0,156	0,121	0,099
sulphates	-0,017	-0,036	0,062	-0,027	0,017	0,059	0,135	0,074	0,156	1,000	-0,017	0,054
alcohol	-0,121	0,068	-0,076	-0,451	-0,360	-0,250	-0,449	-0,780	0,121	-0,017	1,000	0,436
quality	-0,114	-0,195	-0,009	-0,098	-0,210	0,008	-0,175	-0,307	0,099	0,054	0,436	1,000

Tabulka 3 – Korelační matice hodnot proměnných s barevně škálovým zvýrazněním (upraveno v Excelu).



Čtvercová korelační matice obsahuje vzájemné **korelace**³ jednotlivých proměnných a je symetrická podle hlavní diagonály. Díky škálovému barevnému zvýraznění jsou lépe patrné nižší i vyšší hodnoty, díky čemuž z ní lze na první pohled předběžně vyčíst například celkem silnou přímou korelaci (0,839) mezi hustotou a zbytkovým cukrem (čím více cukru, tím větší hustota) a naopak třeba nepřímou korelaci (-0,780) mezi hustotou a alkoholem (čím více alkoholu, tím nižší hustota). Její grafická podoba (Graf 3) pak zobrazuje totéž v matici jednoduchých dvourozměrných bodových grafů.



Graf 3 – Grafická korelační matice.

³ korelace – souvztažnost, vzájemný vztah ⁽¹⁾

Vysoké hodnoty korelačních koeficientů vysvětlujících proměnných mohou svědčit o **multikolinearitě**. Tuto domněnku podporuje i determinant korelační matice (pouze části vysvětlujících proměnných, tedy bez posledního řádku a sloupce proměnné *quality*), který má hodnotu $|R| = 0,00622$, tedy velmi blízkou nule. Nejmenší charakteristické číslo korelační matice vysvětlujících proměnných je $\lambda_{\min} = 0,0206$ (více než o řád nižší než druhé nejmenší), což je sice značně malá hodnota, avšak na prokazatelný signál multikolinearity ještě dostatečně malou hodnotou⁴ není. Největší charakteristické číslo pak je $\lambda_{\max} = 3,2223$, tudíž odmocnina poměru těchto dvou hodnot $\sqrt{\lambda_{\max}/\lambda_{\min}} = 12,492$, čili ani tento indikátor multikolinearity zcela nepotvrzuje (výsledek je menší než 30).

Hodnoty VIF⁵ překračují hraniční hodnotu 5 pouze u třech proměnných (viz Tabulka 4). Nejvyšších hodnot těchto ukazatelů dosahuje proměnná *density* (hustota). Ta již dle korelační matice (Tabulka 3) silně korelovala s proměnnými *resid_sugar* (zbytkový cukr +0,839) a *alkohol* (alkohol -0,780), které obě právě i zde výrazně naznačují výskyt multikolinearity. Z fyzikálního hlediska je přitom vztah těchto proměnných zřejmý: cukry zvyšuje hustotu, kdežto řídký alkohol ji naopak snižuje. Jelikož tyto dvě proměnné přímo „regulují“ hodnotu hustoty, zkusme tuto proměnnou z modelu vypustit a znovu prověřit charakteristiky multikolinearity.

Proměnná	R_j		VIF _j
fixed_acid	0,793	✓	2,691
volatile_acid	0,352	✓	1,141
citric_acid	0,377	✓	1,165
resid_sugar	0,960	✗	12,644
chlorides	0,438	✓	1,237
free_sdo	0,664	✓	1,788
total_sdo	0,744	✓	2,239
density	0,982	✗	28,233
pH	0,738	✓	2,196
sulphates	0,349	✓	1,139
alcohol	0,933	!	7,707

Tabulka 4 – Vícenásobné korelační koeficienty a hodnoty VIF vysvětlujících proměnných.

Proměnná	R_j		VIF _j
fixed_acid	0,512	✓	1,356
volatile_acid	0,337	✓	1,128
citric_acid	0,371	✓	1,160
resid_sugar	0,551	✓	1,435
chlorides	0,411	✓	1,204
free_sdo	0,653	✓	1,745
total_sdo	0,732	✓	2,153
pH	0,499	✓	1,331
sulphates	0,232	✓	1,057
alcohol	0,627	✓	1,647

Tabulka 5 – Viz Tabulka 4, při vynechání proměnné *density* (hustota). (Upraveno v Excelu.)

Z nově vypočtených hodnot VIF a vícenásobných korelačních koeficientů (viz Tabulka 5) je patrné, že po vypuštění proměnné *density* jsou již hodnoty všech ostatních proměnných v normě. Determinant korelační matice zbylých vysvětlujících proměnných je nyní $|R| = 0,17565$, tedy již výrazně vzdálenější nule. Odmocnina poměru největšího a nejmenšího charakteristického čísla je $\sqrt{\lambda_{\max}/\lambda_{\min}} = \sqrt{2,457/0,287} = 2,9266$, čili opět vhodnější hodnoty.

Pro eliminaci multikolinearity tedy proměnnou *density* z regresní analýzy vynecháme.

⁴ Za významně malé charakteristické číslo se považují hodnoty s první platnou číslicí na třetím nebo dalším desetinném místě. ⁽³⁾

⁵ VIF – Variance Inflation Factors – hodnoty ležící na hlavní diagonále inverzní korelační matice vysvětlujících proměnných.

Vícerozměrná regresní analýza

Pro volbu relevantních parametrů ze zbylých vysvětlujících proměnných byla použita metod hledání nejlepšího modelu postupným přidáváním proměnných (**forward**). Ta do výběru v devíti krocích postupně zařadila všechny proměnné, kromě `citric_acid` (kyselina citrónová), která nevyhověla při sekvenčních F-testech.

```
Výsledky- vícerozm. regrese(Krok 9, konečné řešení)
již žádné F na zahrnutí nepřesahuje daná meze
Záv.prom.: quality vícenás. R = ,52223865 F = 203,6728
R2= ,27273321 sv = 9,4888
Poč. případů: 4898 upravené R2= ,27139414 p = 0,000000
Směrodatná chyba odhadu: ,755967031
Abs.člen: 2,061453482 Sm. chyba: ,3481241 t(4888) = 5,9216 p = ,0000

alcohol b*=-,504 volatile_acid b*=-,22 resid_sugar b*=-,147
free_sdo b*=-,091 fixed_acid b*=-,05 sulphates b*=-,054
total_sdo b*=-,04 pH b*=-,028 chlorides b*=-,02
```

Výstup 1 – Výstupní hlášení vícerozměrné regresní analýzy.

Při použití standardní metody byla tatáž proměnná označena jako pro model nevýznamná a její hodnota pro individuální **t-testy** o regresních parametrech ve své absolutní hodnotě (-0,3011) výrazně nedosahovala ani potřebného tabulkového minima 1,6449, tj. kritéria t-testu 95% kvantilu při 4888 stupních volnosti ⁽²⁾.

	b*	Sm.chyba - z b*	b	Sm.chyba - z b	t(4888)	p-hodn.
Abs. člen			2,06145	0,348124	5,9216	0,000000
alcohol	0,503625	0,015612	0,36244	0,011235	32,2594	0,000000
volatile_acid	-0,222225	0,012773	-1,95260	0,112234	-17,3976	0,000000
resid_sugar	0,146619	0,014591	0,02560	0,002548	10,0489	0,000000
free_sdo	0,091418	0,016106	0,00476	0,000839	5,6761	0,000000
fixed_acid	-0,048975	0,013788	-0,05140	0,014470	-3,5520	0,000386
sulphates	0,053806	0,012515	0,41754	0,097117	4,2994	0,000017
total_sdo	-0,042030	0,017871	-0,00088	0,000372	-2,3518	0,018722
pH	0,028380	0,014053	0,16645	0,082424	2,0195	0,043494
chlorides	-0,023981	0,013286	-0,97211	0,538558	-1,8050	0,071132

Tabulka 6 – Výsledky regrese se závislou proměnnou `quality`.

Testové kritérium F pro celkový **F-test** je 203,6728. Tato hodnota bohatě dostačuje na zamítnutí nulové hypotézy, tedy že žádná z uvažovaných vysvětlujících proměnných do regresní rovnice nepatří ⁽³⁾, neboť v 95% kvantilu F-rozdělení je hodnota testového kritéria F při 9 a 4888 stupních volnosti dle tabulek ⁽²⁾ je 1,8799. Hodnota parametru $p = 0,000000$ ve výstupu přímo ukazuje, že pravděpodobnost chybného zamítnutí této nulové hypotézy je menší (a při větší přesnosti výpočtu ještě o několik řádů menší) než 0,0001%.

Index determinace je zde značen jako R^2 (R^2) má hodnotu 0,273 a **index korelace** (R) pak 0,522, což sice značí významnou závislost vysvětlované proměnné na vysvětlujících proměnných (je významný od nuly), bohužel ne však zrovna nejvyšší model (do celé jedničky má daleko).

```
quality = 2,06145
+ 0,36244 * alcohol
- 1,95260 * volatile_acid
+ 0,02560 * resid_sugar
+ 0,00476 * free_sdo
- 0,05140 * fixed_acid
+ 0,41754 * sulphates
- 0,00088 * total_sdo
+ 0,16645 * pH
- 0,97211 * chlorides
```

Výstup 2 – Výsledná regresní rovnice.

Z výsledné regresní rovnice (Výstup 2) je patrné že nejvlivnější proměnnou je *volatile_acid* (obsah těkavých kyselin), která se dle popisných statistik (Tabulka 2) pohybuje pouze v rozmezí 0,08 – 1,1. Podle této rovnice tedy platí „čím méně těkavých kyselin, tím vyšší kvalita“. To samé platí i o proměnné *chlorides* (chloridy), i když s cca polovičním vlivem, a také s menší možností rozsahu nastavení (0,009 – 0,346). Záporně ovlivňují kvalitu ještě proměnné *fixed_acid* a *total_sdo*, ostatní proměnné ji ovlivňují kladně, tzn. větší hodnota znamená větší kvalitu (např. *sulphates*, *alcohol*, *pH* apod).

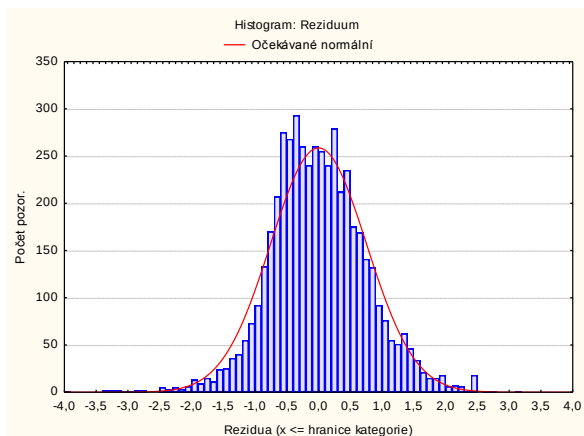
Testy reziduí

Pro testy reziduí byla pro každé pozorování pomocí výsledné regresní rovnice (Výstup 2) vypočtena predikovaná hodnota závislé proměnné (*quality*) a rozdíl mezi ní a hodnotou při měření zjištěnou (pozorovanou) pak tvoří reziduum (odchylku, resp. rušivou složku).

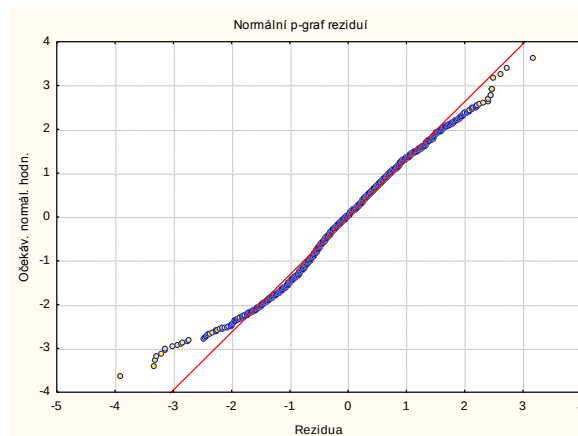
	Pozorovaná	Predikovaná	Reziduum
N platných	4898	4898	4898
Průměr	5,87791	5,87791	0,00000
Medián	6,00000	5,86747	-0,03256
Součet	28 790,00	28 790,00	0,00
Minimum	3,00000	3,98539	-3,90468
Maximum	9,00000	7,24695	3,17732
Rozptyl	0,78436	0,21392	0,57044
Sm. odch.	0,88564	0,46251	0,75527
Šikmost	0,15580	-0,06101	0,06582
Špičatost	0,21653	-0,10765	1,07052

Tabulka 7 – Popisné statistiky pozorované a predikované hodnoty závislé proměnné a odchylek mezi nimi.

Základní popisné statistiky (Tabulka 7) jsou tedy vypočteny z jednotlivých tří sloupců hodnot, tedy pozorované kvality, predikované kvality a rezidua (rozdílu mezi nimi).

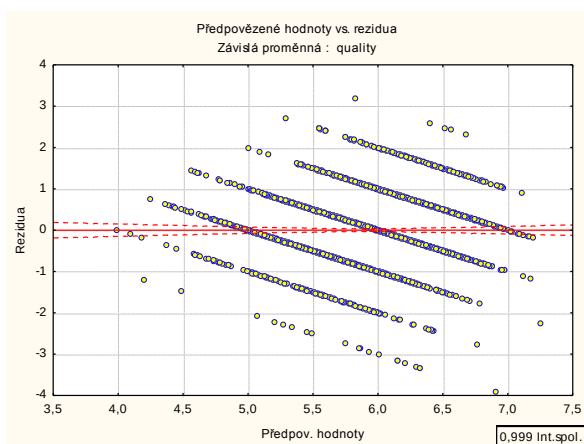


Graf 4 – Rozdělení reziduí.

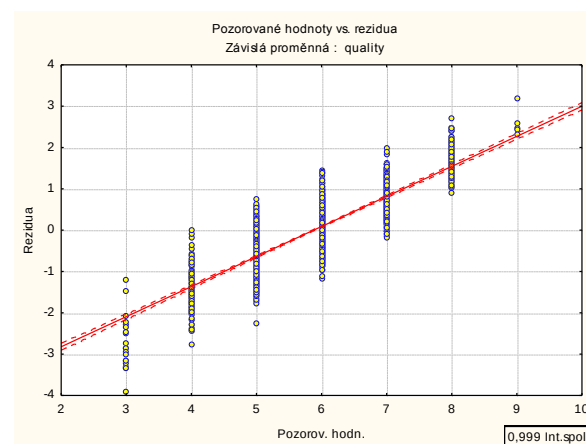


Graf 5 – Normální pravděpodobnostní graf reziduí.

Histogram reziduí (Graf 4) až na drobné odchylky nasvědčuje jejich normálnímu rozdělení. Totéž značí i normální pravděpodobnostní graf (Graf 5), v němž by jednotlivé body v případě normálního rozdělení měly ležet na stejné přímce (vyznačena červeně). Toto se zdá být, opět vyjma krajních případů, dodrženo.



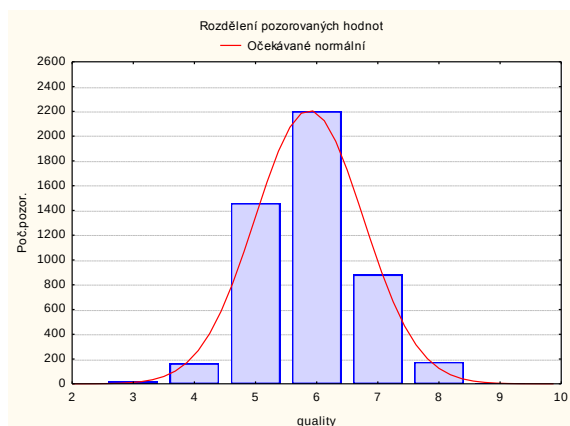
Graf 6 – Rezidua (osa Y) v závislosti na predikovaných hodnotách (osa X).



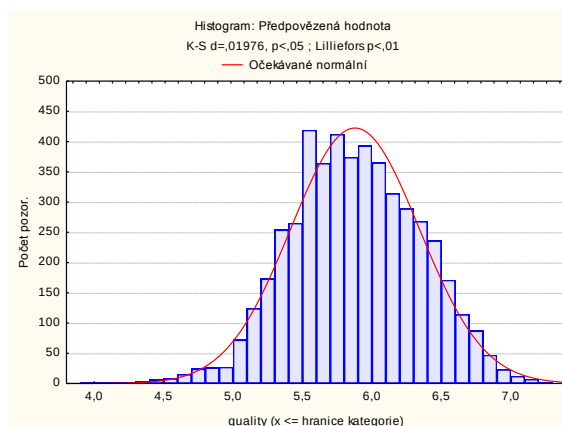
Graf 7 – Rezidua v závislosti na pozorovaných hodnotách.

Z grafu reziduí v závislosti na predikovaných hodnotách (Graf 6) je již patrná určitá nevěrohodnost modelu, ovšem je třeba brát u tohoto grafu v potaz, že z něho nelze vyčíst hustotu jednotlivých bodů, která je v místech okolo 0 na ose Y nejvyšší. I tak je ovšem zjevné, že mimo pravděpodobnostní pás, byť 99,9%, leží podstatná část bodů a to dosti výrazně.

Graf 7 zobrazuje rezidua v závislosti na pozorovaných hodnotách. Zde je jasné vidět, že pro nižší pozorované závislé hodnoty jsou odchylky záporné a pro vyšší naopak kladné. To svědčí o nutnosti dorovnávat tyto krajní hodnoty rušivou složkou, čili model je v těchto krajích ještě méně spolehlivý než ve středu, což je okolo hodnoty 6.

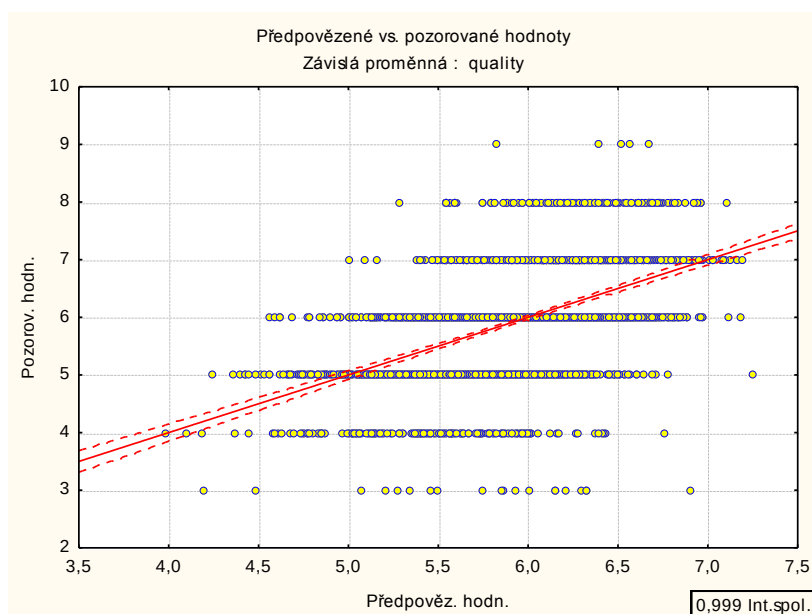


Graf 8 – Rozdělení pozorovaných hodnot.



Graf 9 – Rozdělení predikovaných hodnot.

Pro porovnání jsou uvedeny histogramy pozorovaných (Graf 8) a predikovaných (Graf 9) hodnot závislé proměnné (kvality). Pozorované hodnoty byly pouze celočíselné, proto nelze vytvořit jemnější histogram. Oproti tomu predikované hodnoty jsou spojité, načež byly rozděleny na více rozsahových skupin a to po 0,1. I zde je patrné, že krajní hodnoty nejsou téměř pokryty a většina predikcí směřuje do rozmezí 5 – 7.



Graf 10 – Pozorované a predikované hodnoty.

Bodový Graf 10 na ose Y zobrazuje pozorované hodnoty a na ose X hodnoty predikované. V ideálním případě by veškeré body měly ležet na přímce procházející hlavní diagonálou kvadrantu (červená přímka), případně jí být co nejbližší. Vzhledem k faktu, že pozorované hodnoty proměnné *quality* jsou celočíselné, dochází k jejich skokovým rozestupům na ose Y. Bohužel, jak již bylo řečeno dříve, spousta predikovaných hodnot je velmi vzdálena odhadovanému ideálu a výslednou regresní rovnici je tudíž třeba brát s rezervou.

Na tomto i předchozích grafech jsou zároveň patrná i některá **vlivná pozorování**. Ta se samozřejmě krom odlehlosti bodů v grafu dají identifikovat i dle generovaných statistik. Kromě pozorované a predikované hodnoty a rezidua program nabízí i několik

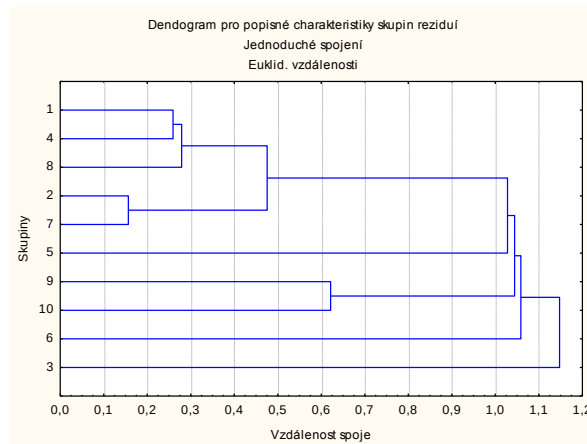
údajů pro identifikaci odlehlých pozorování. Konkrétně jde například o standardizované předpovědi, standardizované residua, Cookovu vzdálenost a Mahalanobisovu vzdálenost. Všechny tyto charakteristiky identifikovaly tytéž vlivné body, z nichž nejodlehlejší bylo pozorování č. 4746, kvalita pozorovaná 3, predikovaná 6,905, residuum -3,905. Např. Cookova vzdálenost zde dosáhla hodnoty 0,1592, což je o řád asi vyšší než druhý nejvyšší případ (0,0199), Mahalanobisova vzdálenost byla 260,7427 a standardizované residuum -5,165. Druhým nejvlivnějším bodem pak bylo pozorování č. 775 s pozorovanou kvalitou 9 a predikovanou 5,823. Oba tyto body (a i další) lze v grafu (např. Graf 10) dle těchto souřadnic snadno pozorovat.

Pro test **heteroskedasticity** byla zjištěná residua náhodně roztržena do deseti stejně velkých skupin (viz Tabulka 8). Nad nimi byly postupně provedeny Bartlettův a Levenův test. Výsledkem Bartlettova testu byla hodnota $b = 13,002$, přičemž 95% kvantil rozdělení chí-kvadrát s 9 (M-1) stupni volnosti $v_{0,95}(9) = 16,919$. Levenův test skončil s výsledkem testového kritéria $f = 1,0172$, a 95% kvantil rozdělení F při 9 (M-1) a 4888 (n-M) stupních volnosti je dle tabulek ⁽²⁾ rovna hodnotě $f_{0,95}(9,4888) = 1,8799$. Do ní se vypočtené F také s rezervou vejde. Z obou testů vyplývá, že hypotézu o rovnosti rozptylů reziduí v jednotlivých skupinách nelze na 5% hladině významnosti zamítnout a regresní model je tudíž homoskedastický, čili z tohoto hlediska nezávadný.

Skupina	Počet	Průměr	Medián	Minimum	Maximum	Rozptyl	Sm.odch.	Šikmost	Špičatost
1	490	-0,0012	-0,0290	-3,3211	2,6038	0,6518	0,8073	-0,1071	1,1829
2	490	-0,0404	-0,0538	-2,8540	2,4039	0,4840	0,6957	0,0158	1,1488
3	490	0,0324	-0,0534	-1,9809	2,4558	0,5669	0,7529	0,4346	0,0526
4	490	-0,0133	-0,0204	-3,2922	2,4558	0,5733	0,7572	-0,0149	1,0183
5	490	0,0021	-0,0509	-2,1175	3,1773	0,5937	0,7705	0,3912	0,9324
6	490	0,0156	-0,0277	-3,9047	2,4558	0,5651	0,7517	-0,0703	2,2711
7	490	-0,0416	-0,0590	-2,7595	2,4558	0,5395	0,7345	0,1003	1,1144
8	490	0,0182	0,0037	-3,3277	2,4827	0,5609	0,7489	-0,0139	1,3863
9	489	0,0339	-0,0094	-2,8646	2,4558	0,5668	0,7529	0,1070	0,5755
10	489	-0,0056	-0,0077	-3,1500	2,2510	0,6067	0,7789	-0,1771	0,9978

Tabulka 8 – Popisné statistiky skupin reziduí, nad kterými byl proveden test heteroskedasticity.

Vytvoříme-li pro tyto popisné statistiky jednotlivých náhodně zvolených skupin reziduí **dendrogram**⁶, je hned zjevné, které skupiny a jak moc výrazně se odlišují od ostatních (Graf 11). V tomto případě například nejvíce mimo stojí skupiny 3 a 6, kdežto nejbližše si jsou (do Euklidovské vzdálenosti 0,5) skupiny 1, 4, 8, 2 a 7. Skupinové rozptyly naštěstí zásadní rozdíly nevykazují.



Graf 11 – Dendrogram pro popisné statistiky (Tabulka 8) skupin reziduí.

⁶ dendrogram – diagram pro znázornění vzdáleností skupin při shlukové analýze

Závěr

Vícerozměrný regresní model, popisující kvalitu vína v závislosti na jeho jednotlivých složkách, se podařilo úspěšně sestavit (viz Výstup 2). Ač tato výsledná regresní rovnice vyhověla testovaným kritériím platnosti, z grafů je patrné, že přesnost tohoto modelu není zrovna stoprocentní. O tom ostatně svědčí i index korelace této rovnice o hodnotě pouhých 0,522. Ač tedy model pro automatizovaný výpočet kvality vína přímo využít nelze, ukazuje nám jeho rovnice proměnné, které výslednou kvalitu vína ovlivňují nejvíce a které naopak nejméně, či dokonce takřka vůbec, tj. proměnné nezařazené do modelu nebo odvozené z hodnot jiných proměnných. Podle něho tedy například kvalita vína pomůže nižší obsah těkavých kyselin a chloridu, naopak zvýšené hodnoty parametrů typu sírany, alkohol či pH kvalitu zvyšují.

Jak již bylo zmíněno v úvodu, ve zdrojových datech byla zastoupena především normální vína, tudíž ta kvalitní a nekvalitní se v nich vyskytla jen zřídka. I z těchto důvodů model lépe predikuje hodnoty vedoucí ke střední kvalitě (`quality = 6`) a okrajové případy (0-4 a 8-10) téměř nepostihuje.

Program použitý pro provedené analýzy – Statistica 9 – se během práce s ním ukázal jako velmi uživatelsky přívětivý a obsahující značné možnosti statistických analýz. Pro běžné používání při zpracování analýz v rámci různých vědeckých prací jej lze rozhodně doporučit.

Literatura

1. **Cortez, Paulo.** Wine Quality Data Set. *UC Irvine Machine Learning Repository*. [Online] Center for Machine Learning and Intelligent Systems, 10 7, 2009. [Cited: 4 1, 2010.] <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>.
2. **Hebák, Petr a Kahounová, Jana.** *Počet pravděpodobnosti v příkladech*. Praha : Informatorium, 1994. 80-85427-48-6.
3. **Hebák, Petr a kolektiv.** *Vícerozměrné statistické metody [2]*. Praha : Informatorium, 2005. 80-7333-036-9.
4. **Tvrdlík, Josef.** Analýza vícerozměrných dat. *Ostravská univerzita - Doktorské studium*. [Online] Ostravská univerzita v Ostravě, Přírodovědecká fakulta, 2003. [Citace: 25. 2 2010.] http://prf.osu.cz/doktorske_studium/dokumenty/Multivariable_Data_Analysis.pdf.