Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra matematiky

Bakalářská Práce

Úspěšnost tenistů na grandslamových turnajích

Plzeň 2013 Puhlovský Vladimír

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW, atd.) uvedené v přiloženém seznamu.

V Plzni dne 26.5.2013 ……..

podpis

Poděkování

Velmi rád bych poděkoval vedoucímu mojí bakalářské práce Mgr. Michalu Frieslovi, Ph.D., za jeho odborné vedení, vstřícný přístup a podnětné připomínky během vytváření této práce.

Abstrakt

Tématem této práce je statistická analýza závislostí získaných dat. V práci jsou zformulovány hypotézy, které je třeba otestovat. Je využíváno párového t-testu a nástrojů regresní analýzy, u které jsou ověřovány její předpoklady. Nulová střední hodnota reziduí je testována t-testem, konstantnost rozptylu pomocí Spearmanova testu. Pro odstranění heteroskedasticity je použita metoda vážených nejmenších čtverců.

Klíčová slova: Párový t-test, regresní analýza, heteroskedasticita, t-test, Spearmanův test, metoda vážených nejmenších čtverců.

Abstract

The subject of this thesis is a statistical analysis of dependence of collected data. In the thesis are formulated hypothesis, which need to be tested. There is used pair t-test and tools of regression analysis, in which are verified their presumptions. Null mean value of residuals is tested by the t-test, permanency of the variance is tested by the Spearman test. For elimination of heteroskedasticity is used the method of weighted least squares.

Key words: Pair t-test, regression analysis, heteroskedasticity, t-test, Spearman test, weighted least squares method

Obsah

Seznam tabulek

1 Úvod

Cílem bakalářské práce je otestovat hypotézy, týkající se úspěšnosti tenistů na grandslamových turnajích za období 2000-2011. Začátek práce je zaměřen na popis jednotlivých grandslamových turnajů. V další kapitole je krátce popsán software, který bude používán. V následující kapitole je pak ukázáno, jak a odkud byla data získána, popřípadě jakým způsobem jsme hodnotili úspěšnost jednotlivých tenistů.

V kapitole č. 5 je uvedena teorie, která bude posléze aplikována na námi získaná data. Následuje základní statistické zpracování bodové úspěšnosti tenistů a poté jsou zformulovány hypotézy, které je potřeba vyšetřit. První hypotéza se týká srovnání jednotlivých povrchů. To znamená, jestli v jednotlivých letech existuje statisticky významný rozdíl ve výkonech tenistů, vždy na dvojici turnajů. To samé bude provedeno pro vybrané národnosti. Druhá hypotéza se zaměřuje na to, zda existuje vliv národnosti tenistů, to v jaké ruce drží raketu a jejich výšky na jejich průměrnou úspěšnost za dané období. To samé budeme zjišťovat na jednotlivých turnajích.

2 O grandslamových turnajích

Grand Slam, v počeštěné podobě také často grandslam, je nejvyšší kategorie mužského a ženského profesionálního okruhu v tenise, která v jedné sezóně zahrnuje čtyři turnaje – nazývané také jako „majory“. V současné podobě jsou to tyto turnaje

- Australian Open (tvrdý povrch)
- French Open (antuka)
- Wimbledon (tráva)
- US Open (tvrdý povrch).

V původním smyslu bylo za Grand Slam označováno vítězství ve všech čtyřech nejvýznamnějších mezinárodních tenisových turnajích během jedné kalendářní sezóny, tzv. „čistý grandslam“ a od roku 1983 po rozhodnutí Mezinárodní tenisové federace i v řadě za sebou během dvou sezón, tzv. „nekalendářní grandslam“. V přeneseném významu jsou pak také tyto čtyři turnaje označovány jako grandslam nebo grandslamové turnaje. Získání Grand Slamu je vzácné – například v mužské dvouhře se to za celou otevřenou éru podařilo pouze Rodu Laverovi v roce 1969. Poslední ženou, která získala Grand Slam ve dvouhře, byla Steffi Grafová v roce 1988. Vzhledem ke stále větší specializaci hráčů (především v mužské kategorii) a k faktu, že jednotlivé grandslamové turnaje se hrají na odlišném povrchu (především French Open se svým pomalým antukovým povrchem liší od všech ostatních), je dnes získání Grand Slamu ve dvouhře čím dál méně pravděpodobné. Výjimkou z pravidla se stal v letech 2006 a 2007 Roger Federer, kterému Grand Slam těsně unikl – vyhrál Wimbledon, Australian Open a US Open, na French Open ale prohrál ve finále s Rafaelem Nadalem. V roce 2010 získal tři tituly Rafael Nadal, ale skrečoval čtvrtfinále čtvrtého Australian Open. V sezóně 2011 pak zvítězil na třech grandslamech Novak Djoković, který prohrál v semifinále French Open.[12]

2.1 Australian Open

Obrázek 1: Australian Open

Autralian Open, The Grand Slam of Asia/Pacific, je jeden ze čtyř tenisových turnajů nejvyšší kategorie Grand Slamu, každoročně hraný v druhé polovině ledna a jediný na jižní polokouli. Úvodní ročník se konal v roce 1905 jako amatérské mezinárodní mistrovství Australásie a do roku 1927 nesl název Australasian Championships. Poté, až do počátku otevřené éry tenisu v roce 1968 se jmenoval Australian Championships. V sezóně 1969 byl turnaj poprvé otevřen profesionálním tenistům, což odráží část „Open" v názvu. Od roku 1972 je dějištěm grandslamu opět Melbourne. Z původního areálu Kooyong Lawn Tennis Club s travnatými dvorci se v roce 1988 událost přemístila do dnešního Melbourne Parku, v němž jsou kurty se středně tvrdým akrylátovým povrchem Plexicushion Prestige s lepší konzistencí a nižším zadržováním tepla. Ten byl položen roku 2008, do té doby se hrálo na povrchu Rebound Ace. Roger Federer a Serena Williamsová jsou jedinými hráči, kteří zvítězili na obou površích. Dva hlavní dvorce Rod Laver Arena a Hisense Arena disponují zatahovací střechou a lze je využít při dešti či vysokých teplotách. Každoroční návštěvnost turnaje je vysoká. V roce 2010 dosáhla 24hodinová návštěvnost rekordní výše v historii všech grandslamů, když za jediný den do areálu přišlo 77 043 diváků, celkový úhrn ročníku 2010 pak činil 653 860 návštěvníků. [12]

2.2 French Open

Obrázek 2: French Open

French Open, známý pod názvem Roland Garros (oficiálně francouzsky Les Internationaux de France de Roland Garros nebo Tournoi de Roland-Garros, také Mezinárodní mistrovství Francie v tenise), je druhý ze čtyř tenisových turnajů nejvyšší kategorie Grand Slamu, každoročně hraný na přelomu května a června. Pojmenování získal po francouzském válečném letci Rolandu Garrosovi, který roku 1913 jako první přeletěl Středozemní moře. Úvodní ročník se konal v roce 1891 pod názvem Championnat de France. Až do počátku otevřené éry tenisu v roce 1968

se v angličtině jmenoval French Championships. V sezóně 1968 byl turnaj poprvé otevřen profesionálním tenistům, což odráží část „Open“ v názvu. Od roku 1928 je dějištěm události osm hektarů rozlehlý tenisový areál Stade Roland Garros, který se nachází na území 16. Městského obvodu Paříže v sousedství Boulogneského lesíka. Je to jediný ze čtyř nejvýznamnějších turnajů velké čtyřky hraný na otevřených antukových dvorcích. Největším z dvaceti kurtů je stadión Court Philippe Chatrier s kapacitou převyšující 15 tisíc diváků, jenž získal název po francouzském sportovním funkcionáři Philippu Chatrierovi. Druhý velký dvorec pro deset tisíc osob nese jméno Court Suzanne Lenglenpo ženské tenisové legendě Suzanne Lenglenové Grandslam je nejsledovanější francouzskou událostí ve světě. Pro pomalý antukový povrch, sedm kol a absenci tiebreaku v rozhodující sadě utkání, které muži hrají na tři vítězné sety, je French Open považován za fyzicky nejnáročnější tenisový turnaj světa. Na počest čtyř francouzských tenistů první poloviny 20. Století označovaných jako „čtyři mušketýři“ získává vítěz mužské dvouhry Pohár mušketýrů. [12]

2.3 Wimbledon

The Championships Wimbledon, nebo pouze Wimbledon, je nejstarší a nejslavnější tenisový turnaj na světě. Představuje třetí ze čtyř událostí nejvyšší kategorie – Grand Slamu, každoročně hranou na přelomu června a července v jihozápadní londýnské části Wimbledon. Úvodní ročník se na travnatých dvorcích All England Clubu uskutečnil v roce 1877. Od sezóny 1995 je tráva stříhána na výšku 8 milimetrů a pro svou odolnost je od roku 2001 používán 100 % víceletý jílek vytrvalý. Největším z celkového počtu 49 kurtů, z nichž 19 je určeno pro grandslam, se nazývá jednoduše centrální dvorec. Ten má od sezóny 2009 zatahovací střechu pro případ deště v typicky proměnlivém jihoanglickém počasí. Důležitou roli ve Wimbledonu hraje tradice a historie. Šatny jsou původní dřevěné. Areál, včetně dvorců, neobsahuje žádné reklamní plochy a tenisté musí hrát v předepsaném oblečení, které je bílé. Typické je podávání jahod se šlehačkou a šampaňského. V neděli předělující první a druhý týden se tradičně nehraje. Patronát nad

Obrázek 3: Wimbledon

grandslamem drží britská královská rodina v čele s patronkou klubu anglickou královnou Alžbětou II. A jeho prezidentem, kterým je od roku 1969 její bratranec vévoda z Kentu, jmenovitě Jeho královská výsost Princ Edward. Ten spolu s chotí předává ceny vítězům. Od roku 1887 získává vítěz mužské dvouhry pozlacený stříbrný pohár, který nahradil původní trofej „Field Cup“ (1877–1883), respektive později tři roky udělovaný „Challenge Cup“ (1884–1886). [12]

2.4 US Open

Obrázek 4: US Open

United States Open (častěji U.S. Open či US Open, Mezinárodní mistrovství USA v tenise) je jeden z nejprestižnějších mezinárodních tenisových turnajů, každoročně uzavírá grandslamovou sérii turnajů. Koná se na přelomu srpna a září v Národním tenisovém centru Billie Jean Kingové. Od roku 1978 se koná v New Yorku na tvrdém povrchu DecoTurf. Do roku 1974 se hrálo na trávě a v letech 1975 až 1977 na antuce. Centrální kurt Arthur Ashe Stadium je pojmenován po Arthurovi Ashovi, afroamerickém tenisovém hráči, který vyhrál úvodní mužské finále v otevřené éře v roce 1968. Všechny dvorce jsou osvětlené. V roce 2005 byly kurty turnajů série US OPEN přemalovány – vnitřní plocha hřiště získala modrou a vnější zelenou barvu. Výsledným efektem je zvýšení viditelnosti míče. [12]

3 Použitý software

Pro zpracování dat byl vybrán statistický software Matlab od společnosti The Math-works. Matlab je integrované prostředí pro vědecko-technické výpočty, modelování, návrhy algoritmů, simulace, analýzu a prezentaci dat, paralelní výpočty, měření a zpracování signálů, návrhy řídicích a komunikačních systémů. Matlab je nástroj jak pro pohodlnou interaktivní práci, tak pro vývoj širokého spektra aplikací.[13]

Pro učely bakalářské práce byl využíván Statistics Toolbox, který nabízí rozsáhlý soubor nástrojů pro práci s daty. Zahrnuje funkce a interaktivní nástroje pro modelování dat, analýzu trendů, simulaci stochastických systémů a vývoj algoritmů pro statistiku. Statistics Toolbox podporuje širokou škálu úloh od výpočtů základní popisné statistiky až po vývoj a vizualizaci mnohorozměrných nelineárních modelů. Dále nabízí velké množství statistických grafů a interaktivních grafických nástrojů, jako je polynomiální prokládání a modelování výsledkových ploch.[14]

Část práce týkající se srovnání jednotlivých povrchů byla provedena v MS Office Excel.

4 Shromáždění a zpracování dat

Data byla shromažďována z různých tenisových portálů, jako je www.tenisportal.cz a z oficiálních webů jednotlivých turnajů. Jedná se o dosažené výsledky tenistů na turnajích (tzn.: v jaké fáze turnaje tenista skončil) a informace o tenistech, jako je národnost tenisty, v jaké ruce drží raketu a jeho výška. Tato data byla shromážděna za období 2000 až 2011.

4.1 Hrací systém turnajů a bodové hodnocení výkonů

Všechny grandslamové turnaje se hrají stejným systémem. Tenisté jsou rozlosováni do pavouka a dále hrají vyřazovacím způsobem, a to následovně: 1. kolo, 2. kolo, 3. kolo, osmifinále, čtvrtfinále, semifinále, finále. Každého turnaje se zúčastní 128 tenistů. Úspěšnost tenistů jsme hodnotili bodově podle toho, v jaké části turnaje vypadl.

	Dosažený výsledek	Bodové ohodnocení
1.	Kolo	1 bod
2.	Kolo	2 body
3.	Kolo	3 body
	Osmifinále	4 body
	Čtvrtfinále	5 bodů
	Semifinále	6 bodů
	Poražený finalista	7 bodů
	Vítěz	8 bodů

Tabulka 1: Bodovací systém výsledků

Dále jsme zjišťovali výšku tenisty, to zda drží raketu v pravé nebo levé ruce a jeho národnost. Soubor dat čítá 569 tenistů, kteří se ve zkoumaném období zúčastnili alespoň jednoho turnaje. V následující tabulce je znázorněn vzorek získaných dat. V tabulce (2) ukazujeme výsledky z prvních a posledních dvou turnajů za námi zkoumané období. Pomlčka ve výsledcích znamená, že se tenista turnaje nezúčastnil.

				2000		2011	
Jméno	Výška v cm	Hraje rukou	Národnost	AO	FO	W	US
Roger FEDERER	185	P	SUI	3	4	5	6
Leyton HEWITT	180	P	AUS	4	4	2	-
Rafael NADAL	185	L	ESP	-	-	7	7
Andy RODDICK	187	P	USA	-	-	3	5
Juan – Carlos FERRERO	182	P	ESP	6	3	-	4
Novak DJOKOVIČ	187	P	SER	-	-	8	8
Tommy ROBREDO	180	P	ESP	-	-	1	-
David NALBANDIAN	180	P	ARG	-	-	3	3
Marat SAFIN	193	P	RUS	1	5	-	-
Nikolaj DAVYDENKO	177	P	RUS	-	-	1	3
Michajl JUZNYJ	182	P	RUS	-	-	4	1
Fernando GONZALEZ	182	P	CHI	-	-	3	1
Andre AGASSI	180	P	USA	8	2	-	-
David FERRER	175	P	ESP	-	-	4	4

Tabulka 2: Vzorek získaných dat

5 Základní statistické pojmy

Pro další zpracování uvedeme některé základní pojmy, které budeme dále využívat. První část této kapitoly je zpracována podle publikací [1] a [4]. Teoretická část týkající se regresní analýzy, pak podle publikace [1] a [3].

Aritmetický průměr souboru $x_1, x_2, \ldots, x_n$ je hodnota

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,.$$

Jsou-li v_i $(i = 1, 2, \ldots, n)$ nějaká reálná čísla, pak součet $\sum_{i=1}^{n} x_i \; v_i$ se nazývá *vážený průměr* hodnot $x_1, x_2, \ldots, x_n$ s vahami $v_1, v_2, \ldots, v_n$.

Rozptyl statistického souboru $x_1, x_2, \ldots, x_n$ je hodnota

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Číslo $\sigma_n = \sqrt{\sigma_n^2}$ se nazývá *směrodatná odchylka*.

Pro $n > 1$ se definuje též *výběrový rozptyl*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

a *výběrová směrodatná odchylka* $s = \sqrt{s^2}$.

Korelační koeficient Buď (X, Y) dvourozměrná veličina s kladnými rozptyly $D(X), D(Y)$. Korelační koeficient $\varrho = \varrho(X, Y)$ je definován předpisem

$$\varrho = \frac{cov(X, Y)}{\sqrt{D(X)D(Y)}}$$

a je mírou statistické lineární závislosti veličin X, Y.

Normální rozdělení $N(\mu, \sigma^2)$ s parametry $\mu, \sigma > 0$ je rozdělení pravděpodobnosti určené hustotou

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in (-\infty, \infty).$$

Rozdělení $N(0,1)$ s parametry $\mu = 0, \sigma^2 = 1$ se nazývá *normované normální rozdělení.* Její hustota je pak

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad x \in (-\infty, \infty),$$

a distribuční funkce rozdělení $N(0,1)$ je definována vztahem

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt, \qquad x \in (-\infty, \infty).$$

p-hodnota testu Při vyhodnocování testů se často využívá tzv. *p-hodnota* příslušné statistiky. Platí, že uvažovanou hypotézu lze zamítnout na hladině významnosti $\boldsymbol{\alpha}$, právě tehdy když p-hodnota příslušné testové statistiky je menší nebo rovna $\boldsymbol{\alpha}$.

5.1 Testy normality

χ^2 – test dobré shody Mějme náhodný výběr rozsahu n z rozdělení nějaké náhodné veličiny X, která může být spojitá nebo diskrétní. Chceme na hladině významnosti α testovat hypotézu, že rozdělení veličiny X se rovná nějakému rozdělení, které známe až na hodnotu m neznámých parametrů (může být $m = 0$, známe-li všechny parametry). Test, který zde popíšeme, se nazývá χ^2 – *test dobré shody*.

Postupujeme v následujících krocích:

1. Rozdělíme obor hodnot na několik nepřekrývajících se tříd, jejich počet označíme k.
2. Zjistíme, kolik hodnot realizovaného náhodného výběru se nachází v jednotlivých třídách; tyto počty označíme $n_i (i = 1, 2, \dots, k)$.
3. Neznáme-li některé parametry předpokládaného modelu (tj. $m > 0$) odhadujeme je.
4. Pro každou třídu spočteme očekávaný počet hodnot v této (řekněme i-té) třídě, $o_i = np_i$ $(i = 1, 2, \dots, k)$, kde n je rozsah výběru a p_i je pravděpodobnost, že

veličina X s předpokládaným rozdělením pravděpodobnosti nabude hodnoty patřící do i-té třídy. Čísla o_i jsou obecně necelá, jejich součet je n.

5. Je-li některý očekávaný počet o_i menší než 5, sdružíme danou třídu s některou jinou třídou. Sdružená třída má očekávaný počet o_i roven součtu očekávaných počtů ze tříd, jejichž sdružením vznikla. Ve sdružování pokračujeme, dokud není pro každou třídu $o_i \geq 5$; počet nových tříd označíme opět k .
6. Hypotézu, že veličina se řídí předpokládaným modelem, zamítáme na hladině významnosti α, je-li

$$\sum_{i=1}^{k} \frac{(n_i - o_i)^2}{o_i} > \chi^2_{1-\alpha}(v),$$

kde $v = k - 1 - m$ (předpokládá se $v > 0$, tj. počet tříd $k > m + 1$).

Lillieforsův test Tento test lze provádět i pro náhodné výběry poměrně malých rozsahů. Nechť x_i $(i = 1, 2, \dots, n)$ je uspořádaný náhodný výběr. *Výběrová (emperická) distribuční funkce (EDF)* F_n je definována vztahem

$$F_n(x) = \frac{počet, kolik\ hodnot\ ve\ výběru\ je \leq x}{n}.$$

Funkce $F_n(x)$ je tedy po částech konstantní (schodovitá) s hodnotami

$$F_n(x) \begin{cases} 0 \quad pro \quad x < x_1 \\ \frac{i}{n} \quad pro \quad x \in [x_i, x_{i+1}) \; i = 1, 2, \dots, n-1 \\ 1 \quad pro \quad x \geq x_n \end{cases}$$

Lillieforsův test dobré shody je založen na statistice D, která je definována jako největší vzdálenost mezi hodnotami výběrové distribuční funkce a hypotetické distribuční funkce, $D = sup\{|F_n(x) - F(x)| : x \in R\}$. Hodnotu D lze snadno získat ze vzorce

$D = max\{D^+, D^-\}$, kde

$D^+ = max\left\{\frac{i}{n} - F(x_i) : i = 1, \dots, n\right\}$,

$$D^- = max\left\{F(x_i) - \frac{i-1}{n} : i = 1, \dots, n\right\}.$$

Ekvivalentní vyjádření statistiky D je

$$D = \frac{1}{2n} + max\left\{\left|\frac{i-0{,}5}{n} - F(x_i)\right| : i = 1, \dots, n\right\}.$$

Hypotézu, že náhodný výběr pochází z rozdělení s distribuční funkcí F, zamítáme v případě, že statistika D nabude vysoké hodnoty.

Popsaný test se v případě, že hypotetická distribuční funkce F veličiny spojitého typu je předem zcela určena včetně hodnot parametrů, tj. k určení parametrů hypotetického rozdělení se nepoužívají hodnoty náhodného výběru, nazývá *Kolmogorovův test.* Výše definované statistiky D^+, D^- lze použít k jednostranným testům, jsou rovněž uvedeny pro $1 \leq n \leq 1000$ a vybraná α kritické hodnoty $D^+_{1-\alpha}$ statistiky D^+ takové, že za platnosti hypotézy platí $P(D^+ > D^+_{1-\alpha}(n)) = \alpha$. Pro Kolmogorovovu statistiku D pak lze použít příbuzný vztah $P\left(D > D^+_{1-\alpha}(n)\right) = 2\alpha$, přičemž pro $\alpha \leq 0{,}1$ je aproximace velmi přesná pro libovolné n. Pro velké n lze pro běžně používané hladiny významnosti α použít statistiku D přibližnou kritickou hodnotu $\sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}$.

Jestliže parametry hypotetického rozdělení neznáme a odhadujeme je pomocí hodnot náhodného výběru, jsou kritické hodnoty statistiky D jiné než hodnoty tabelované pro (standardní) Kolmogorovův test s předem určeným rozdělením. Uvažujeme např. test dobré shody s normálním rozdělením $N(\mu, \sigma^2)$, kde jsou parametry μ, σ^2 neznámé. Odhadneme je výběrovými hodnotami $\bar{x} = \frac{1}{n}\sum x_i$, $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$, a distribuční funkci $F(x) = \Phi(\frac{x-\mu}{\sigma})$ odhadneme pomocí $\Phi(\frac{x-\bar{x}}{s})$. Test založený na statistice D se pak nazývá *Lillieforsův test.* Pro velké n lze pro $\alpha = 0{,}05$ (resp. $\alpha = 0{,}01$) použít přibližnou kritickou hodnotu $0{,}89/\sqrt{n}$ (resp. $1{,}04/\sqrt{n}$).

Jarque-Berův test Jedná se o test založený na koeficientech šikmosti a špičatosti. Hodnotu testové statistiky určíme podle vztahu

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

kde

n je počet dat,

$S = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^{\frac{3}{2}}}$ je koeficient šikmosti,

$K = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}$ je koeficient špičatosti.

Testová statistika má při platnosti nulové hypotézy χ^2 rozdělení se dvěma stupni volnosti. Nulovou hypotézu zamítáme v případě

$$JB > \chi^2_{1-\alpha}(v = 2).$$

5.2 Testy o shodnosti středních hodnot

Párový t-test Uvažujeme dvourozměrnou normální náhodnou veličinu (X, Y) a označíme $\mu_1 = E(X), \mu_2 = E(Y)$. Buď $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ realizace náhodného výběru z rozdělení této veličiny a chtějme na hladině významnosti α testovat hypotézu $H_0 : \mu_1 = \mu_2$. V tomto případě nelze obecně předpokládat vzájemnou nezávislost náhodných výběrů $x_1, x_2, \ldots, x_n$ a $y_1, y_2, \ldots, y_n$, neboť hodnota y_i může nějakým způsobem souviset s x_i. Následující postup se nazývá *t-test pro závislé výběry* neboli *párový t-test*. Označme:

$$z_i = y_i - x_i \qquad \text{pro } i = 1, 2, \ldots, n.$$

Tyto hodnoty jsou realizací náhodného výběru z rozdělení veličiny $Z = Y - X$, která má střední hodnotu $\mu = \mu_2 - \mu_1$. Hypotéza $H_0 : \mu_1 = \mu_2$ je pak ekvivalentní rovnosti $\mu = 0$. Testujeme-li např. H_0 proti oboustranné alternativě $\mu_1 \neq \mu_2$ při neznámých rozptylech, použijeme statistiku

$$t = \frac{\bar{z}}{s}\sqrt{n},$$

kde $\bar{z}$ a s je průměr a výběrová směrodatná odchylka souboru $z_1, z_2, \ldots, z_n$, s kritickým oborem $|t| > t_{1-\frac{\alpha}{2}}(n-1)$.

5.3 Regresní analýza

5.3.1 Metoda nejmenších čtverců

Pro zjišťování závislosti budeme využívat metodu nejmenších čtverců, kterou nyní zformulujeme a zároveň zformulujeme předpoklady této metody.

Mějme maticový tvar regresního modelu

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (1)$$

kde

$\boldsymbol{y}$ je vektor hodnot vysvětlované proměnné,

$\boldsymbol{X}$ je matice hodnot vysvětlujících proměnných,

$\boldsymbol{\beta}$ je vektor hledaných regresních koeficientů,

$\boldsymbol{\varepsilon}$ je vektor náhodné složky.

Předpoklady metody nejmenších čtverců jsou

$(P1)\ E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$

$(P2)\ var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$

(P3) silnější podmínka zahrnující předcházející $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n) \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$

(P4) $\boldsymbol{X}$ je nestochastická matice, která má plnou hodnost.

Předpoklad (P1) vyjadřuje to, že náhodná složka modelu $\boldsymbol{\varepsilon}$ je nevychýlená, tedy chyby modelu nemají systematický charakter.

Předpoklad (P2) zachycuje dvě vlastnosti, variabilita náhodné složky je konstantní pro $i = 1, 2, \dots, n$ (homoskedasticita) a jednotlivé prvky náhodné složky jsou nezávislé, $cor(\varepsilon_i, \varepsilon_{i'}) = 0$ pro $i \neq i'$.

Pokud jsou splněny podmínky (P1) a (P2) mluvíme o klasickém lineárním regresním modelu, pokud je navíc splněna podmínka (P3) hovoříme normálním lineárním regresním modelu. Podmínka (P4) vyjadřuje požadavek, že vysvětlující proměnné, které tvoří sloupce matice $\boldsymbol{X}$ jsou navzájem nezávislé. Protože

předpokládáme počet měření, která máme k dispozici (n) je větší než počet proměnných, které odhadujeme (p), tedy $p < n$ je hodnost matice $\boldsymbol{X}$ rovna p.

Je-li dán model ve tvaru (1) a platí podmínky (P1), (P2) a (P4), pak odhad regresních parametrů metodou nejmenších čtverců minimalizující výraz

$$Q(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}).$$

je

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

a platí

$$E(\boldsymbol{b}) = \boldsymbol{\beta}, \quad var(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \quad s^2 = \frac{(\boldsymbol{y} - \boldsymbol{Xb})^T(\boldsymbol{y} - \boldsymbol{Xb})}{n - p},$$

kde s^2 je nestranným odhadem σ^2.

Dále nadefinujeme tzv. *vyrovnané hodnoty* $\hat{y}_i, (i = 1, 2, \dots, n)$, což jsou složky vektoru očekávaných hodnot, vztahem

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb}.$$

Rozdíly mezi naměřenými a vyrovnanými hodnoty se značí $e_i, (i = 1, 2, \dots, n)$, a nazývají se *rezidua:*

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}.$$

5.3.2 Kvalita regresního modelu

K hodnocení kvality regresního modelu využíváme koeficient determinace R^2. ***Koeficient determinace*** definujeme jako hodnotu odvozenou z celkových čtverců následujícím způsobem

$$R^2 = \frac{S_V^2}{S_T^2} = 1 - \frac{RSS}{S_T^2}$$

kde *RSS* je *residuální součet čtverců* ve tvaru

$$RSS = \boldsymbol{e}^T\boldsymbol{e} = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 ,$$

S_V^2 je *vysvětlený součet čtverců* daný tvarem

$$S_V^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

a S_T^2 je *celkový součet čtverců* ve tvaru

$$S_T^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

pokud k odhadům vyrovnaných hodnot použijeme metodu nejmenších čtverců, pak podle Pythagorovy věty platí

$$S_T^2 = RSS + S_V^2, \quad R^2 \in \langle 0,1 \rangle.$$

5.3.3 Heteroskedasticita v regresi

Mějme σ_i^2 rozptyl náhodné složky $\varepsilon_i, (i = 1, \dots, n)$. V homoskedastickém modelu jsou tyto rozptyly stejné, ${\sigma_i}^2 = \sigma^2$. Není-li tato podmínka splněna, říkáme, že jde o model *heteroskdastický*. Použijeme-li v heteroskedastickém modelu standardní metodu nejmenších čtverců, je odhad b vektoru regresních koeficientů β i nadále nestranný, není už však nejlepší ve třídě nestranných lineárních odhadů. Navíc bychom obdrželi nesprávné intervaly a pásy spolehlivosti a testy významnosti by mohly mít větší pravděpodobnost chyby 1. druhu, než kolik bylo požadováno volbou hladiny významnosti.

5.3.3.1 Metody zjišťování heteroskedasticity

Spearmanův test Tento test je klasickým testem konstantnosti rozptylu založeným na pořadových statistikách. Zkoumá korelaci pořadí mezi jednou vysvětlující proměnou a rezidui. Při jeho provádění postupujeme v následujících krocích.

1. Absolutní hodnoty reziduí $|e_i|$ seřadíme vzestupně a očíslujeme.
2. Pořadové číslo přiřadíme k původním (tj. nesrovnaným) reziduím.
3. Absolutní hodnoty vysvětlující proměnné $|x_j|$ seřadíme vzestupně a očíslujeme.
4. Pořadové číslo přiřadíme k původním (tj. nesrovnaným) hodnotám x_j.
5. Spočítáme rozdíly v pořadí reziduí a pozorování: $d_i =$ pořadí $|e_i|$ - pořadí $|x_j|$.
6. Spočítáme Spearmanův koeficient korelace pořadí:

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

kde d_i je diference v pořadí příslušných dvojic. Hodnoty blízké jedničce ukazují na heteroskedasticitu dat. Přesněji odvozená statistika

$$t = r\sqrt{\frac{n-p}{1-r^2}}$$

má při platnosti nulové hypotézy studentovo t-rozdělení s $\nu = n - p$ stupni volnosti.

Dalšími testy pro zjištění heteroskedasticity jsou např.: Goldfeld-Quandtův test, Glejsrův test.

5.3.4 Metoda vážených nejmenších čtverců

Uvažujeme heteroskedastický model ve tvaru

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (1).$$

V takovém případě jsou rozptyly náhodné složky úměrné (až na násobek, který označujeme σ^2) známým hodnotám c_i ($i = 1, 2, \dots, n$), tj. $D(\varepsilon_i) = c_i\sigma^2$ $(i = 1, 2, \dots, n)$.

Označme $w_i = \frac{1}{c_i}$ $(i = 1, 2, \dots, n)$. Předpokládejme tedy, že platí

$$D(\varepsilon_i) = \frac{\sigma^2}{w_i} \quad (i = 1, 2, \dots, n), \quad (2)$$

kde w_i jsou známé konstanty a σ^2 je neznámý rozptyl.

Ukážeme, že heteroskedastický model (1) lze transformovat na model homoskedastický. Označíme $y_i^{(w)} = y_i\sqrt{w_i}$, $\varepsilon_i^{(w)} = \varepsilon_i\sqrt{w_i}$, a $\boldsymbol{y}^{(w)}, \boldsymbol{\varepsilon}^{(w)}$ sloupcové vektory s těmito složkami. Podobně označíme $X^{(w)}$ matici s prvky $x_{ij}\sqrt{w_i}$ $(i = 1, 2, \dots, n\,; j = 1,2, \dots, p)$. Matice $\boldsymbol{X}^{(w)}$ tedy vznikne tak, že i-tý řádek regresní matice $\boldsymbol{X}$ násobíme hodnotou $\sqrt{w_i}$. Zapsáno maticově

$$\boldsymbol{y}^{(w)} = \begin{bmatrix} y_1^{(w)} \\ y_2^{(w)} \\ \vdots \\ y_n^{(w)} \end{bmatrix}, \qquad \boldsymbol{\varepsilon}^{(w)} = \begin{bmatrix} \varepsilon_1^{(w)} \\ \varepsilon_2^{(w)} \\ \vdots \\ \varepsilon_n^{(w)} \end{bmatrix}$$

$$\boldsymbol{X}^{(\boldsymbol{w})} = \begin{bmatrix} x_{11}\sqrt{w_1} & x_{12}\sqrt{w_1} & \cdots & x_{1p}\sqrt{w_1} \\ x_{21}\sqrt{w_2} & x_{22}\sqrt{w_2} & \cdots & x_{2p}\sqrt{w_2} \\ \vdots & \vdots & & \vdots \\ x_{n1}\sqrt{w_n} & x_{n2}\sqrt{w_n} & \cdots & x_{np}\sqrt{w_n} \end{bmatrix}.$$

Dostáváme tedy již homoskedastický model

$$\boldsymbol{y}^{(\boldsymbol{w})} = \boldsymbol{X}^{(\boldsymbol{w})}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{(\boldsymbol{w})}, \quad (4)$$

protože pro rozptyly nových náhodných složek použitím (2) dostaneme

$$D(\varepsilon_i^{(w)}) = D\left(\varepsilon_i\sqrt{w_i}\right) = w_i D(\varepsilon_i) = \sigma^2 \quad (i = 1, 2, \dots, n).$$

Odhad $\boldsymbol{b}$ vektoru regresních koeficientů $\boldsymbol{\beta}$ modelu (1) stanovíme tak, že použijeme metodu nejmenších čtverců pro homoskedastický model (4).

Pro vyrovnané hodnoty $\hat{\boldsymbol{y}}^{(\boldsymbol{w})}$ modelu (4) platí

$$\hat{\boldsymbol{y}}^{(\boldsymbol{w})} = \boldsymbol{X}^{(\boldsymbol{w})}\boldsymbol{b} = \hat{\boldsymbol{y}}\sqrt{\boldsymbol{w}},$$

kde $\hat{\boldsymbol{y}}^{(\boldsymbol{w})}$ jsou vyrovnané hodnoty v modelu (4) a $\hat{\boldsymbol{y}}$ jsou vyrovnané hodnoty v modelu (1). Při použití metody nejmenších čtverců v modelu (4) tedy minimalizujeme součet

$$\sum_{i=1}^{n}(\hat{y}_i^{(w)} - y_i^{(w)})^2 = \sum_{i=1}^{n}(\hat{y}_i\sqrt{w_i} - y_i\sqrt{w_i})^2 = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 w_i.$$

Vzhledem k poslednímu tvaru tohoto součtu se říká, že pro model (1) používáme *metodu vážených nejmenších čtverců* s vahami $w_1, \dots, w_n$.

5.3.5 Test významnosti jednotlivých koeficientů

Uvažujeme regresní model ve tvaru (1). Testujeme významnost jednotlivých regresních koeficientů, tj. hypotézu

$$H_0: \beta_i = 0,$$

Testovací statistika má pak tvar

$$t = \frac{b_i}{s_{b_i}},$$

kde b_i je odhad parametru β_i a $s_{b_i}=\sqrt{s_{b_i}^2}$ je odhad směrodatné odchylky parametru β_i, kde $s_{b_i}^2$ jsou diagonální prvky matice $s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, má rozdělení pravděpodobnosti $t(v=n-p)$. Nulovou hypotézu zamítáme na hladině významnosti α, pokud $|\,t\,| > t_{1-\frac{\alpha}{2}}(n-p)$.

5.3.6 Test významnosti více koeficientů

Uvažujeme opět regresní model ve tvaru (1). Buď $m < p$ a chceme testovat hypotézu, že m zvolených regresních parametrů jsou nulové, tj. hypotézu

$$H_0\text{: } \beta_i = 0 \text{ pro každé } i \in I\ ,$$

kde I je nějaká pevně zvolená množina m různých přirozených čísel, $I \subset \{1, \dots, p\}$.

Buď RSS reziduální součet čtverců modelu se všemi regresními parametry a označíme RSS_H reziduální součet čtverců v modelu po vynechání regresních parametrů s indexy $i \in I$, tj. regrese spočtené za předpokladu platnosti hypotézy H_0. Testovací statistika ve tvaru

$$F = \frac{RSS_H - RSS}{RSS} \cdot \frac{n-p}{m}$$

má rozdělení pravděpodobnosti $F(m, n-p)$. Hypotézu H_0 zamítáme na hladině významnosti α, je-li $F > F_{1-\alpha}(m, n-p)$.

6 Jednoduchá statistická analýza

V následujících tabulkách najdeme přehled základních statistik bodového zisku jednotlivých tenistů za období 2000-2011, jak celkově tak na jednotlivých turnajích. Vzhledem k velkému počtu tenistů uvádíme vždy pouze nejlepších deset tenistů z hlediska celkového součtu bodů.

	Celkem							
	Jméno	Σ	n	$\bar{x}$	σ^2	σ	min	max
1.	Roger FEDERER	278	48	5.79	4.87	2.21	1	8
2.	Lleyton HEWITT	175	44	3.98	3.39	1.84	1	8
3.	Rafael NADAL	173	30	5.77	4.45	2.11	2	8
4.	Andy RODDICK	167	42	3.98	3.88	1.97	1	8
5.	Juan-Carlos FERRERO	140	42	3.33	3.08	1.75	1	8
6.	Novak DJOKOVIČ	139	28	4.96	4.25	2.06	1	8
7.	Tommy ROBREDO	122	42	2.90	1.90	1.38	1	5
8.	David NALBANDIAN	118	33	3.58	2.43	1.56	1	7
9.	Marat SAFIN	118	35	3.37	4.29	2.07	1	8
10.	Nikolaj DAVYDĚNKO	117	41	2.85	2.91	1.70	1	6

Tabulka 3: Základní statistiky – Celkem

Australian Open								
	Jméno	Σ	n	$\bar{x}$	σ^2	σ	min	max
1.	Roger FEDERER	71	12	5.92	3.58	1.89	3	8
2.	Andy RODDICK	47	10	4.70	1.81	1.35	2	6
3.	Lleyton HEWITT	38	12	3.17	2.81	1.67	1	7
4.	Marat SAFIN	38	9	4.22	5.51	2.35	1	8
5.	Rafael NADAL	36	7	5.14	2.12	1.46	3	8
6.	Andre AGASSI	35	5	7.00	1.60	1.26	5	8
7.	Sebastian GROSJEAN	35	10	3.50	2.45	1.57	1	6
8.	David NALBANDIAN	34	9	3.78	2.17	1.47	2	6
9.	Novak DJOKOVIČ	32	7	4.57	7.10	2.66	1	8
10.	Nikolaj DAVYDĚNKO	31	10	3.10	3.09	1.76	1	5

Tabulka 4: Základní statistiky – Australian Open

French Open								
	Jméno	Σ	n	$\bar{x}$	σ^2	σ	min	max
1.	Roger FEDERER	61	12	5.08	5.24	2.29	1	8
2.	Rafael NADAL	52	7	7.43	1.96	1.40	4	8
3.	Juan-Carlos FERRERO	44	11	4.00	4.91	2.22	1	8
4.	Tommy ROBREDO	39	10	3.90	1.49	1.22	1	5
5.	Lleyton HEWITT	38	10	3.80	0.56	0.75	3	5
6.	Nikolaj DAVYDĚNKO	34	10	3.40	2.24	1.80	1	6
7.	Novak DJOKOVIČ	33	7	4.71	2.20	1.48	2	6
8.	Gustavo KUERTEN	31	7	4.43	7.10	2.66	1	8
9.	Fernando GONZALEZ	30	10	3.00	2.80	1.67	1	6
10.	David FERRER	30	9	3.33	1.11	1.05	2	5

Tabulka 5: Základní statistiky – French Open

Wimbledon								
	Jméno	Σ	n	$\bar{x}$	σ^2	σ	min	max
1.	Roger FEDERER	72	12	6.00	6.50	2.55	1	8
2.	Andy RODDICK	50	11	4.55	3.34	1.83	2	7
3.	Lleyton HEWITT	49	12	4.08	3.74	1.93	1	8
4.	Rafael NADAL	42	7	6.00	5.14	2.27	2	8
5.	Novak DJOKOVIČ	34	7	4.86	3.55	1.88	2	8
6.	Feliciano LOPEZ	34	10	3.40	2.04	1.43	1	5
7.	Michajl JUŽNYJ	33	11	3.00	1.45	1.21	1	4
8.	Tim HENMAN	33	8	4.13	3.11	1.76	2	6
9.	Juan-Carlos FERRERO	32	10	3.20	1.56	1.25	1	5
10.	Tomáš BERDYCH	31	8	3.88	2.61	1.62	1	7

Tabulka 6: Základní statistiky – Wimbledon

US Open								
	Jméno	**Σ**	n	$\bar{x}$	σ^2	σ	min	max
1.	Roger FEDERER	74	12	6.17	3.47	1.86	3	8
2.	Andy RODDICK	52	12	4.33	4.39	2.09	1	8
3.	Lleyton HEWITT	50	10	5.00	4.60	2.14	1	8
4.	Rafael NADAL	43	9	4.78	4.17	2.04	2	8
5.	Novak DJOKOVIČ	40	7	5.71	3.35	1.83	3	8
6.	Tommy HASS	36	10	3.60	1.24	1.11	2	5
7.	Tommy ROBREDO	35	10	3.50	0.85	0.92	1	4
8.	Nikolay DAVYDĚNKO	35	11	3.18	2.51	1.59	1	6
9.	André AGASSI	35	7	5.00	3.14	1.77	2	7
10.	Juan-Carlos FERRERO	34	11	3.09	2.63	1.62	1	7

Tabulka 7: Základní statistiky – US Open

Vysvětlivky:

Σ – součet bodů, n – počet turnajů, $\bar{x}$– průměr, σ^2- rozptyl, σ – směrodatná odchylka, min – minimum, max – maximum.

Z hlediska součtu bodů byl nejlepším tenistou Roger Federer, a to jak celkově tak na všech dílčích turnajích. Zároveň se jako jediný tenista za období 2000-2011 zúčastnil všech grandslamových turnajů. Plnou účast na jednotlivých turnajích si připsali pouze Lleyton Hewitt, kterému se to podařilo na dvou turnajích, a to na Australian Open a Wimbledonu a Andy Roddick na jeho domácím turnaji US Open.

Co se týká průměrného počtu bodů, se o nejlepší výkony podělili dva tenisté. Roger Federer, který byl nejlepší celkově, na Australian Open a US Open. Rafael Nadal byl nejlepší na French Open. Ve Wimbledonu dosáhli tyto dva tenisté totožného průměrného výsledku a o první pozici se dělí.

Velmi malý rozptyl ve výkonech zaznamenal Lleyton Hewitt na turnaji French Open a Tommy Robredo na turnaji US Open. Což značí vyrovnanost jejich výkonů na těchto turnajích. Naopak velmi vysoký rozptyl měli výkony Novaka Djokoviče a Marata Safina na turnaji Australian Open, Gustava Kuertena na turnaji French Open a Rogera Federera ve Wimbledonu. To může mít několik příčin, a to nevyrovnanost jejich výkonů, zlepšující se výkony nebo naopak zhoršující se výkony během zkoumaného období.

7 Srovnání jednotlivých povrchů

V této kapitole je naším cílem zjistit, zda existuje statisticky významný vliv povrchu na výkony tenistů z hlediska jejich bodového zisku na jednotlivých turnajích a výkony vybraných národností jako celku v jednotlivých letech. Vždy byly porovnávány dvojice turnajů, čili dvojice rozdílných povrchů.

1. Australian Open – French Open (Akrylát – Antuka)
2. Australian Open – Wimbledon (Akrylát – Tráva)
3. Australian Open – US Open (Akrylát – Beton)
4. French Open – Wimbledon (Antuka – Tráva)
5. French Open – US Open (Antuka – Beton)
6. Wimbledon – US Open (Tráva – Beton)

V každém roce byli vybráni ti tenisté, kteří se zúčastnili dané dvojice turnajů. Pro takto vytvořené výběry byl formulován párový t-test pro dvojici výsledků vybraných tenistů.

Předpokladem párového t-testu je nezávislost jednotlivých dvojic výsledků $(x_1, y_1), \dots, (x_n, y_n)$. V našem případě byl tento předpoklad porušen z toho hlediska, že jednotlivé výsledky jsou na sobě závislé. A to z toho důvodu, že výsledky tenistů jsou omezené počtem účastníků v dané fázi. To znamená, že pokud by se dané dvojice turnajů zúčastnili všichni tenisté (tzn. 128 tenistů), tak by si střední hodnoty takovýchto výběrů byly rovny a takovýto test ztrácí význam. V naší práci nikdy k případu, že by se dané dvojice turnajů zúčastnilo všech 128 tenistů, nedochází, avšak stále to je velký počet tenistů, proto testům všech tenistů dohromady nepřikládáme velkou váhu. Daleko větší význam pro nás má testování jednotlivých národností, kde už je počet tenistů takový, že závislost mezi jednotlivými dvojicemi výsledků nehraje takovou roli.

Do tabulky (8) byly zaneseny p-hodnoty tohoto testu pro každou dvojici turnajů v jednotlivých letech a počet tenistů, pro které byl párový t-test počítán.

	AO-FO	AO-W	AO-US	FO –W	FO-US	W-US
2000	0.46	0.09	0.48	0.29	0.46	0.43
	(93)	(94)	(87)	(102)	(103)	(101)
2001	0.13	0.12	0.22	0.32	0.50	0.23
	(97)	(93)	(88)	(96)	(97)	(100)
2002	0.39	0.46	0.48	0.22	0.48	0.27
	(98)	(96)	(89)	(94)	(90)	(92)
2003	0.34	0.12	0.26	0.42	0.36	0.43
	(95)	(86)	(85)	(103)	(100)	(98)
2004	0.32	0.25	0.38	0.16	0.35	0.26
	(100)	(95)	(94)	(99)	(96)	(93)
2005	0.50	0.50	0.35	0.50	0.45	0.40
	(101)	(96)	(92)	(100)	(97)	(101)
2006	0.15	0.34	0.24	0.39	0.40	0.47
	(96)	(92)	(86)	(101)	(95)	(101)
2007	0.45	0.44	0.47	0.34	0.32	0.47
	(94)	(96)	(97)	(101)	(95)	(100)
2008	0.34	0.45	0.47	0.45	0.24	0.45
	(92)	(95)	(87)	(97)	(93)	(94)
2009	0.50	0.07	0.47	0.29	0.16	0.13
	(97)	(93)	(91)	(100)	(96)	(94)
2010	0.23	0.45	0.36	0.42	0.44	0.50
	(96)	(94)	(92)	(102)	(95)	(100)
2011	0.44	0.38	0.47	0.36	0.30	0.44
	(103)	(102)	(90)	(101)	(95)	(99)

Tabulka 8: p-hodnoty párového t-testu – Celkem

Pokud porovnáme p-hodnoty testů s hladinou významnosti $\alpha = 0.05$, tak vidíme, že pokud testujeme všechny tenisty, tak se nám nikde nepodařilo prokázat statisticky významný rozdíl v jejich výkonech mezi jednotlivými turnaji. Nyní však z těchto výběrů vybereme určitou národnost a postup budeme opakovat. To si ukážeme na příkladu Španělska, které jsme vybrali z důvodu většího počtu tenistů (v závorce je vždy uveden jejich počet).

	AO-FO	AO-W	AO-US	FO –W	FO-US	W-US
2000	0.033	0.367	0.400	0.281	0.086	0.500
	(11)	(7)	(8)	(8)	(9)	(6)
2001	0.072	0.255	0.355	0.003	0.107	0.068
	(14)	(15)	(15)	(14)	(16)	(15)
2002	0.064	0.010	0.321	0.064	0.015	0.221
	(12)	(8)	(10)	(10)	(11)	(8)
2003	0.064	0.173	0.179	0.053	0.011	0.500
	(12)	(9)	(12)	(9)	(12)	(11)
2004	0.274	0.218	0.322	0.017	0.069	0.378
	(14)	(13)	(11)	(15)	(11)	(11)
2005	0.037	0.404	0.329	0.161	0.067	0.381
	(15)	(12)	(12)	(13)	(13)	(11)
2006	0.500	0.299	0.380	0.169	0.054	0.153
	(10)	(8)	(9)	(12)	(12)	(11)
2007	0.061	0.432	0.172	0.047	0.210	0.276
	(10)	(10)	(10)	(13)	(12)	(12)
2008	0.218	0.294	0.411	0.375	0.027	0.181
	(12)	(11)	(11)	(11)	(12)	(10)
2009	0.291	0.344	0.500	0.410	0.169	0.500
	(14)	(13)	(13)	(13)	(13)	(12)
2010	0.066	0.434	0.003	0.073	0.129	0.004
	(11)	(11)	(10)	(13)	(11)	(11)
2011	0.429	0.204	0.288	0.429	0.276	0.399
	(11)	(12)	(10)	(11)	(11)	(10)

Tabulka 9: p-hodnoty párového t-testu pro národnost Španělsko

Pokud porovnáme p-hodnoty testů z tabulky (9) s hladinou významnosti $\boldsymbol{\alpha = 0.05}$, tak vidíme, že některé p-hodnoty nám ukazují na statisticky významný rozdíl mezi dvěma turnaji. Většinou se jedná o rozdíl mezi turnajem French Open (antuka) a zbytkem turnajů. Lze tedy říci, že Španělé podávají rozdílné výkony, pokud hrají na antuce. Na příkladu Španělů jsme tedy ukázali, že lze sledovat rozdílné výkony mezi dvojicemi turnajů, pokud se zaměříme na jednotlivé národnosti, což může být dáno například historickou zvyklostí v daných zemích.

8 Vliv národnosti, ruky a výšky celkově a na jednotlivých turnajích

V této kapitole se zaměříme na testování vlivu vysvětlujících proměnných „národnost“, „ruka“(držení rakety) a „výška“ na průměrný výkon tenisty za období 2000-2011 na grandslamových turnajích. Pro testování těchto vlivů jsme zformulovali následující model

$$\overline{y_{ijp}} = \mu + \alpha_i + \beta_j + \gamma\left(x_{ijp} - \bar{x}\right) + \varepsilon_{ijp} \quad (1)$$

$$(p = 1, \dots, n_{ij};\ i = 1,2, \dots, I;\ j = 1,2)$$

kde $\alpha_I = 0$, $\beta_2 = 0$,

$\overline{y_{ijp}}$ $\left(p = 1, \dots, n_{ij}\right)$ jsou průměrné výkony tenistů ve třídě, která je kombinací i-té úrovně vysvětlující proměnné „národnost“ a j-té úrovně vysvětlující proměnné „ruka“ a n_{ij} je počet pozorování pro tuto kombinaci (může být též $n_{ij} = 0$).

α_i jsou regresní parametry odpovídající vlivu i-té úrovně vysvětlující proměnné „národnost“ na výkon tenisty, sloupce regresní matice odpovídající parametrů α_i mají na i −tém řádku 1 pokud tenista náleží i-té národnosti nebo 0 pokud ji nenáleží.

β_j je regresní parametr odpovídající vlivu j-té úrovně vysvětlující proměnné „ruka“ na výkon tenisty, sloupec matice odpovídající parametru β_j má na i −tém řádku 0, pokud tenista drží raketu v levé ruce nebo 1 pokud tenista drží raketu v pravé ruce,

γ je regresní parametr odpovídající vlivu výšky na výkon tenisty,

x_{ijp} $\left(p = 1, \dots, n_{ij}\right)$ jsou výšky tenistů ve třídě, která je kombinací i-té úrovně vysvětlující proměnné „národnost“ a j-té úrovně vysvětlující proměnné „ruka“ a n_{ij} je počet pozorování pro tuto kombinaci.

$\bar{x}$ je průměrná výška tenistů,

ε_{ijp} je náhodná složka.

V maticovém zápisu $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ má regresní matice $\boldsymbol{X}$ tvar

$$\boldsymbol{X} = \begin{matrix} n_1 \\ \\ \\ \\ \\ \\ \\ \\ \\ n_{I-1} \\ \\ \\ n_I \\ \\ \end{matrix} \begin{bmatrix} 1 & 1 & & 0 & 1 & x_{111} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ \vdots & \vdots & & & 1 & x_{11n_{11}} \\ & \vdots & & & 0 & x_{121} \\ & \vdots & & & \vdots & \vdots \\ & 1 & & & 0 & x_{12n_{12}} \\ & 0 & & & \uparrow & \uparrow \\ & \vdots & \cdots & \vdots & Ruka & Výška \\ & & & 0 & \downarrow & \downarrow \\ & & & 1 & & \\ & & & \vdots & & \\ & & & 1 & & \\ \vdots & & & 0 & & \\ \vdots & \vdots & & \vdots & & \\ 1 & 0 & & 0 & & \end{bmatrix},$$

kde první sloupec matice $\boldsymbol{X}$ je jednotkový, n_I je počet tenistů I-té národnosti, sloupec označený jako Ruka má na i-tém řádku matice 0, pokud tenista drží raketu v levé ruce nebo 1 pokud drží raketu v pravé ruce a v posledním sloupci jsou výšky jednotlivých tenistů. Vektor regresních parametrů $\boldsymbol{\beta}$ je roven

$$\boldsymbol{\beta} = [\mu, \alpha_1, \dots, \alpha_{I-1}, \beta_1, \gamma]^T.$$

Pro odhad závislosti průměrného výkonu tenisty na uvedených vlivech použijeme metodu nejmenších čtverců. Použitím průměrného výkonu tenisty, však do modelu zanášíme heteroskedasticitu. V našem případě je tedy rozptyl náhodné složky $var(\varepsilon_{ijp}) = c_{ijp}\sigma^2$. Pro odstranění heteroskedasticity použijeme metodu vážených nejmenších čtverců a naším cílem je tedy určit konstantu c_{ijp}.

Mějme tedy výsledky v_i od jednoho tenisty, kde $i = (1, \dots, k)$ a k je počet výsledků, z nichž počítáme průměrný výkon daného tenisty. Jelikož počítáme průměr z výsledků od jednoho tenisty, jedná se tedy o průměr ze závislých pozorování. Rozptyl náhodné složky ε_i je roven rozptylu průměrného výkonu tenisty. Rozptyl takového průměru má následující tvar

$$D(\frac{1}{k}\sum_{i=1}^{k} v_i) = \frac{1}{k^2}\sum_{i=1}^{k} D(v_i) \; =$$

$$= \frac{1}{k^2}\left(\sum_{i=1}^{k} D(v_i) \; + \; 2\sum_{i=1}^{k}\sum_{\substack{j=2 \\ i<j}}^{k} cov(v_i, v_j)\right) = \frac{1}{k^2}\left(k\sigma^2 + \; 2\sum_{i=1}^{k}\sum_{\substack{j=2 \\ i<j}}^{k} cov(v_i, v_j)\right) =$$

$$= \frac{\sigma^2}{k} + \frac{2\sum_{i=1}^{k}\sum_{\substack{j=2 \\ i<j}}^{k} cov(v_i, v_j)}{k^2} = \frac{\sigma^2}{k} + \frac{\frac{k!}{(k-2)!}.\hat{\varrho}_C.\sigma^2}{k^2} \; = c.\sigma^2,$$

kde $\frac{k!}{(k-2)!}$ je počet dvojic (v_i, v_j).

Z poslední rovnosti vyplývá, že tvar konstanty c je

$$c = \frac{\frac{\sigma^2}{k} + \frac{\frac{k!}{(k-2)!}.\hat{\varrho}_C.\sigma^2}{k^2}}{\sigma^2} = \frac{\frac{\sigma^2.(k+k(k-1).\hat{\varrho}_C)}{k^2}}{\sigma^2} = \frac{1 + (k-1).\hat{\varrho}_C}{k}.$$

Takto určíme konstantu pro každého tenistu v modelu a budeme ji značit c_{ijp}.

Kovarianci (resp. přímo korelační koeficient) jsme odhadovali tak, že jsme v každé skupině výsledků se stejnými úrovněmi vysvětlujících proměnných „národnost“ a „ruka“ („výška“ zde nebyla využita z toho důvodu, že by se většinou vytvořily skupiny po jednom tenistovi) od každého tenisty vybrali dvojici výsledků s tím, že informace obsažena v ostatních výsledcích zůstane bohužel nevyužita. Byly vybrány vždy první a poslední dvojice výsledků a z nich byl korelační koeficient odhadnut. Z výpočtů byli vyřazeni ti tenisté, kteří zaznamenali pouze jeden výsledek nebo celé skupiny, u nichž nebylo možné korelační koeficient určit. Poté jsme vzali odhady z jednotlivých skupin a váženým průměrem určili celkový korelační koeficient. Odhady z jednotlivých skupin jsme vážili počtem tenistů, ze kterých byl korelační koeficient počítán. Celkový korelační koeficient tedy odhadneme jako

$$\hat{\varrho}_C = \frac{\sum_{i=1}^{k} n_i \hat{\varrho}_i}{n},$$

kde

$\hat{\varrho}_i$ je odhadovaný korelační koeficient v jednotlivých skupinách,

$\hat{\varrho}_C$ je odhad celkového korelačního koeficientu,

n_i je počet tenistů v $i-$té skupině (po vyřazení),

n je celkový počet tenistů (po vyřazení).

Dostáváme tedy následující model

$$\overline{y_{ijp}}^{(w)} = \mu\sqrt{w_{ijp}} + \alpha_i{}^{(w)} + \beta_j{}^{(w)} + \gamma\left(x_{ijp} - \bar{x}\right)^{(w)} + \varepsilon_{ijp}{}^{(w)}, \quad (2)$$

kde

$w_{ijp} = \frac{1}{c_{ijp}}$.

Metodou vážených nejmenších čtverců tedy odhadneme regresní parametry modelu a určíme jejich statistickou významnost.

Vliv národnosti na průměrný výkon tenisty testujeme pomocí F-testu významnosti více koeficientů, kterým určíme statistickou významnost koeficientů α_i. Budeme tedy testovat hypotézu, že všechny koeficienty α_i jsou rovny nule.

$$H_0: \alpha_i = 0 \text{ pro každé } i = (1, \dots, I-1),$$

kde I je počet národností v modelu.

Jestli má statisticky významný vliv na průměrný výkon tenisty držení rakety, testujeme pomocí t-testu významnosti regresního koeficientu. V našem případě tedy testujeme hypotézu

$$H_0: \beta_1 = 0,$$

Vliv výšky tenisty na jeho průměrný výkon testujeme opět pomocí t-testu významnosti regresního koeficientu. Testujeme tedy hypotézu

$$H_0: \gamma = 0,$$

zamítnutí námi zformulovaných hypotéz ukazuje na statistický významný vliv některé z vysvětlujících proměnných.

8.1 Regrese celkem za období 2000-2011

V této kapitole budeme zkoumat vliv jednotlivých vysvětlujících proměnných „národnost“, „ruka“, „výška“ za celé období 2000-2011 bez ohledu na to o jaký turnaj se jedná. Odhad významnosti vysvětlující proměnné národnost nám ovlivňuje ten fakt, že některé skupiny národností tvoří malý počet tenistů. Předpokládáme, že nemůžeme objektivně posoudit vliv národnosti na průměrný výkon tenisty, pokud danou skupinu národnosti tvoří malý počet tenistů. Proto jsme skupiny národností, které tvoří méně než 5 tenistů, sloučili do jedné.

Budeme tedy testovat statistickou významnost vysvětlujících proměnných, pro což jsme zformulovali následující hypotézy

- vysvětlující proměnná **„národnost“** :

$$\boldsymbol{H_0}: \alpha_i = 0 \text{ pro každé } i = (1, \dots, I-1),$$

kde I je počet národností v modelu,

- vysvětlující proměnná **„ruka“** :

$$\boldsymbol{H_0}: \beta_1 = 0,$$

- vysvětlující proměnná **„výška“** :

$$\boldsymbol{H_0}: \gamma = 0,$$

Statistickou významnost vysvětlujících proměnných jsme testovali několika přístupy. A to podle toho jakým způsobem jsme odhadovali korelační koeficient (z první nebo poslední dvojice výsledků). A pro porovnání jsme zadali mezní případy kdy korelační koeficient je roven 0 nebo 1. Všechny hypotézy byly testovány na hladině významnosti $\alpha = 0{,}05$.

V tabulce (10) máme uvedeny hodnoty korelačních koeficientů, koeficientů determinace, testovacích statistik a příslušné p – hodnoty námi zformulovaných testů.

	Celkem		Národnost	Ruka	Výška
	$\hat{\varrho}_C$	R^2	F/p	t/p	t/p
první dvojice	0,160	0,504	1,646 0,024	0,252 0,801	2,224 0,027
poslední dvojice	0,379	0,343	1,365 0,108	0,226 0,822	2,401 0,017
	1	0,063	1,123 0,308	0,205 0,840	2,400 0,017
	0	0,725	3,431 4e-8	0,281 0,779	1,242 0,215

Tabulka 10: Výsledky pro celkové hodnocení

Pro odhady korelačního koeficientu z posledních dvojic výsledků a při volbě jedničkového korelačního koeficientu na základě p-hodnoty F-testu významnosti více koeficientů přijímáme hypotézu H_0, to znamená, že vysvětlující proměnná „národnost“ nemá v tomto případě statisticky významný vliv na průměrný výkon tenisty. Při volbě nulového korelačního koeficientu a při odhadu korelačního koeficientu z první dvojice výsledků se naopak jeví statisticky významnou proměnnou. To už se nedá říci o vysvětlující proměnné „ruka“, kde nám p-hodnota t-testu významnosti regresního koeficientu ukazuje na přijmutí hypotézy H_0. Naopak p-hodnoty t-testů významnosti regresního koeficientu vysvětlující proměnné „výška“, které jsou menší než hladina významnosti α, ukazuje na zamítnutí hypotézy H_0 a tedy na statistickou významnost této proměnné u prvních tří modelů.

8.2 Regrese na jednotlivých turnajích za období 2000-2011

Nyní se budeme věnovat každému turnaji zvlášť. Budeme tedy opakovat předchozí postup pro jednotlivé turnaje. To znamená, že pro každý turnaj přepočítáme konstantu c_{ijp} a budeme opět hodnotit vliv vysvětlujících proměnných „národnost“, „ruka“ a „výška“ na Australian Open, French Open, Wimbledonu a US Open opět za celé období 2000-2011.

8.2.1 Australian Open

Prvním z testovaných turnajů je Australian Open, tedy turnaj který se koná na tvrdém akrylátovém povrchu. Alespoň jednoho turnaje se za sledované období zúčastnilo 421 tenistů. V tabulce (11) máme uvedeny hodnoty korelačních koeficientů, koeficientů determinace, testovacích statistik a příslušné p – hodnoty námi zformulovaných testů.

	AO		Národnost	Ruka	Výška
	$\hat{\varrho}_C$	R^2	F/p	t/p	t/p
první dvojice	0,289	0,318	1,491 0,076	-0,160 0,873	0,668 0,504
poslední dvojice	0,313	0,304	1,461 0,087	-0,165 0,869	0,688 0,492
	1	0,065	1,198 0,249	-0,251 0,802	0,920 0,358
	0	0,570	2,700 9e-5	-0,016 0,987	0,149 0,882

Tabulka 11: Výsledky pro turnaj Australian Open

Na základě p-hodnot F-testů významnosti více koeficientů přijímáme hypotézu H_0, to znamená, že vysvětlující proměnná „národnost“ nemá statisticky významný vliv na průměrný výkon tenisty na Australian Open, pro první tři modely. U modelu s volbou nulového korelačního koeficientu se tato proměnná jeví statisticky významnou. O vysvětlující proměnné „ruka“, kde nám p-hodnoty t-testů významnosti regresního koeficientu ukazují na přijmutí hypotézy H_0, se dá říci, že nemá statisticky

významný vliv na průměrný výkon tenisty. P-hodnoty t-testů významnosti regresního koeficientu vysvětlující proměnné „výška“, které jsou větší než hladina významnosti α, ukazují na rozdíl od celkového hodnocení na přijmutí hypotézy H_0 a tedy na statistickou nevýznamnost této proměnné.

8.2.2 French Open

Další testovaným turnajem je French Open, turnaj který se koná na nejpomalejším antukovém povrchu. Alespoň do jednoho turnaje se za sledované období nastoupilo 405 tenistů. V tabulce (12) máme uvedeny hodnoty korelačních koeficientů, koeficientů determinace, testovacích statistik a příslušné p – hodnoty námi zformulovaných testů.

	FO		Národnost	Ruka	Výška
	$\hat{\varrho}_C$	R^2	F/p	t/p	t/p
první dvojice	0,291	0,259	2,986 2e-5	0,097 0,923	2,164 0,031
poslední dvojice	0,227	0,290	3,033 2e-5	-0,021 0,983	2,119 0,035
	1	0,126	2,685 1e-4	0,813 0,417	2,297 0,022
	0	0,481	3,262 3e-6	-0,614 0,540	1,659 0,098

Tabulka 12: Výsledky pro turnaj French Open

Na základě p-hodnot F-testů významnosti více koeficientů zamítáme hypotézu H_0, to znamená, že vysvětlující proměnná „národnost“ má statisticky významný vliv na průměrný výkon tenisty na tomto turnaji. P-hodnoty t-testů významnosti regresního koeficientů ukazují na statistickou významnost tří jednotlivých národnosti, a to Argentiny, Španělska a Ruska. To už se ovšem nedá říci o vysvětlující proměnné „ruka“, kde nám p-hodnoty t-testů významnosti regresního koeficientu ukazují na přijmutí hypotézy H_0. Naopak p-hodnoty t-testů významnosti regresního koeficientu vysvětlující proměnné „výška“ jsou menší než hladina významnosti α, a

ukazují na zamítnutí hypotézy H_0, tedy na statistickou významnost této proměnné, kromě posledního modelu s volbou nulového korelačního koeficientu.

8.2.3 Wimbledon

Třetí testovaný turnaj je londýnský Wimbledon, hraný na travnatém povrchu. Alespoň jeden turnaj za sledované období sehrálo 424 tenistů. V tabulce (13) máme uvedeny hodnoty korelačních koeficientů, koeficientů determinace, testovacích statistik a příslušné p – hodnoty námi zformulovaných testů.

	W		Národnost	Ruka	Výška
	$\hat{\varrho}_C$	R^2	F/p	t/p	t/p
první dvojice	0,254	0,343	1,296 0,168	0,050 0,959	1,459 0,145
poslední dvojice	0,400	0,265	1,159 0,281	0,062 0,950	1,450 0,148
	1	0,060	0,914 0,576	0,049 0,960	1,415 0,158
	0	0,562	2,239 0,001	0,062 0,951	1,488 0,138

Tabulka 13: Výsledky pro turnaj Wimbledon

Na základě p-hodnot F-testů významnosti více koeficientů přijímáme hypotézu H_0, to znamená, že vysvětlující proměnná „národnost“ nemá statisticky významný vliv na průměrný výkon tenisty na Wimbledonu pro první tři modely. U modelu s volbou nulového korelačního koeficientu se tato proměnná jeví statisticky významnou. O vysvětlující proměnné „ruka“, kde nám p-hodnoty t-testů významnosti regresního koeficientu ukazují na přijmutí hypotézy H_0, se dá říci, že nemá statisticky významný vliv na průměrný výkon tenisty. P-hodnoty t-testů významnosti regresního koeficientu vysvětlující proměnné „výška“, které jsou větší než hladina významnosti α, ukazují na přijmutí hypotézy H_0 a tedy na statistickou nevýznamnost této proměnné.

8.2.4 US Open

Posledním testovaným turnajem je americký US Open, hraný na nejrychlejším betonovém povrchu. Do alespoň jednoho turnaje za období 2000-2011 nastoupilo 421 hráčů. V tabulce (14) máme uvedeny hodnoty korelačních koeficientů, koeficientů determinace, testovacích statistik a příslušné p – hodnoty námi zformulovaných testů.

	US		Národnost	Ruka	Výška
	$\hat{\varrho}_C$	R^2	F/p	t/p	t/p
první dvojice	0,317	0,332	1,121 0,246	0,772 0,441	2,143 0,033
poslední dvojice	0,279	0,354	1,239 0,218	0,762 0,447	2,109 0,036
	1	0,065	0,991 0,472	0,880 0,379	2,466 0,014
	0	0,585	2,288 0,001	0,739 0,461	1,577 0,116

Tabulka 14: Výsledky pro turnaj US Open

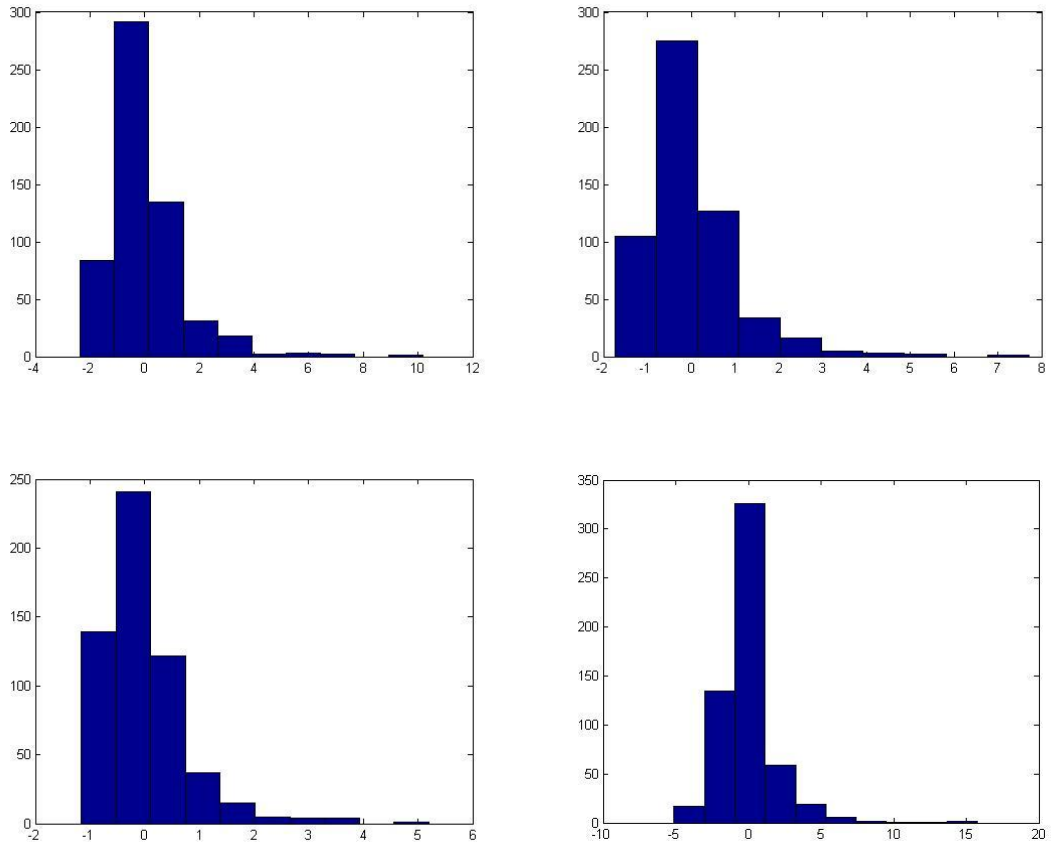
Na základě p-hodnot F-testů významnosti více koeficientů přijímáme hypotézu H_0, to znamená, že vysvětlující proměnná „národnost“ nemá statisticky významný vliv na průměrný výkon tenisty na tomto turnaji. Výjimkou je opět poslední model s volbou nulového korelačního koeficientu. Dále se dá říci o vysvětlující proměnné „ruka“, kde nám p-hodnoty t-testů významnosti regresního koeficientu opět ukazují na přijmutí hypotézy H_0. Naopak p-hodnoty t-testů významnosti regresního koeficientu vysvětlující proměnné „výška“ jsou menší než hladina významnosti α, a ukazují na zamítnutí hypotézy H_0, tedy na statistickou významnost této proměnné, kromě posledního modelu, kde se tato proměnná ukazuje jako statisticky nevýznamná.

8.3 Vyhodnocení modelů

Veškeré používané statistické testy předpokládají, že rezidua pochází z normálního rozdělení. Nyní se tedy budeme zabývat, zda rezidua našich modelů pocházejí z normálního rozdělení. Budeme tedy zkoumat hypotézu

H_0: rezidua pocházejí z normálního rozdělení.

Pro testování normality byly použity testy uvedené v teoretické části: χ^2-test dobré shody, Lillieforsův test a Jarque-Berův test. Tyto testy jsou součástí Statistics Toolbox v Matlabu. Také lze normalitu prvotně prozkoumat pomocí histogramů četností.



Obrázek 5: Histogramy četností pro celkové hodnocení-první dvojice,poslední dvojice,1,0

Už pohled na obrázek (5), na kterém jsou zobrazeny histogramy četností, nám dává najevo, že rezidua modelů pro celkové hodnocení pravděpodobně nepochází z normálního rozdělení. To nám potvrzují i vybrané statistické testy. Kde p-hodnoty testů normality jsou menší než hladina významnosti α=0,05, to znamená, že zamítáme hypotézu H_0 o normalitě reziduí.

Celkem	Lillieforsův test	χ^2-test dobré shody	Jarque-Berův test
	p	p	p
prv. dvojice	<1.000e-3	6.674e-23	<1.000e-3
pos. dvojice	<1.000e-3	1.973e-18	<1.000e-3
1	<1.000e-3	1.608e-12	<1.000e-3
0	<1.000e-3	6.609e-24	<1.000e-3

Tabulka 15: p-hodnoty testů normality reziduí pro celkové hodnocení

Stejná situace se opakuje i na jednotlivých turnajích, tedy ani rezidua modelů pro jednotlivé turnaje nepocházejí z normálního rozdělení. Z testů normality vyplývá, že rezidua všech zkoumaných modelů nepochází z normálního rozdělení. Tento předpoklad statistických testů tedy není splněn. Což výsledky našich testů poněkud oslabuje a nemůžeme se na ně tolik spoléhat.

U všech turnajů můžeme pozorovat, že čím víc se odhad korelačního koeficientu blíží nule, roste koeficient determinace modelu. Při použití jednotkového korelačního koeficientu jsme zaznamenali velmi nízký koeficient determinace. Při použití nulového korelačního koeficientu, však dochází k tomu, že se střední hodnota reziduí nejeví nulová a tudíž jsou modely s nulovým korelačním koeficientem nevhodné. Nulovost střední hodnoty jsme testovali pomocí t-testu. To samé nastává i v případě odhadu korelačního koeficientu z první dvojice výsledků v celkovém hodnocení, kde je tento odhad blízký nule. Pro všechny ostatní odhady je nulovost střední hodnoty reziduí splněna.

Pro detekci heteroskedasticity jsme využívali Spearmanova testu, pro jehož výpočet jsme využili již nadefinované funkce v Matlabu. Ten nám naznačil již homoskedastické modely. Test byl v každém modelu počítán pro každou vysvětlující proměnnou zvlášť.

Na závěr musíme poznamenat, že se ze stejných důvodů jako v předchozí kapitole potýkáme se závislostí mezi jednotlivými výsledky tenistů. A to z toho

důvodu, že výsledky tenistů jsou omezené počtem účastníků v dané fázi turnaje. Což může naše výsledky nějakým způsobem nepříznivě ovlivnit.

8.4 Shrnutí výsledků

První zkoumanou vysvětlující proměnnou byla národnost tenisty. Tato vysvětlující proměnná se ukázala statisticky významnou pouze na turnaji French Open. Můžeme tedy vyslovit následující domněnku, že French Open je díky antukovému povrchu na tolik specifický, že některé národnosti podávají lepší výkony na tomto turnaji a některé naopak. Jak už jsme poznamenali v minulé kapitole 6, projevuje se zde, že v některých zemích je antukový povrch upřednostňován (např.: Španělsko, Argentina). Ve výsledku to znamená, že tyto země mají výhodu oproti zemím, kde se na antukovém povrchu nehraje skoro vůbec a zároveň jim mohou konkurovat na tvrdém a travnatém povrchu, které jsou si svými vlastnostmi podobné. To zřejmě způsobuje ta skutečnost, že antukový povrch je svými vlastnostmi velice obtížný na přizpůsobení. Antukový povrch je mnohem pomalejší, hrají se na něm delší výměny a tím pádem i delší zápasy. Tyto vlastnosti mají národnosti, kde je antukový povrch jakousi tradicí zažité a přechod z rychlého povrchu na ten pomalý antukový a naopak jim nedělá takový problém jako jiným národnostem. V následující tabulce (16) uvedeme hodnoty regresních parametrů „národnost“ modelů, kde se jevila tato vysvětlující proměnná významnou a zároveň byla splněna podmínka nulovosti střední hodnoty náhodné složky a konstantnosti rozptylu náhodné složky.

French Open	První dvojice výsledků	Poslední dvojice výsledků	1
	α_i	α_i	α_i
ARG – Argentina	0,69	0,71	0,57
AUS – Austrálie	0,08	0,10	-0,07
AUT – Rakousko	-0,05	-0,05	-0,06
BEL – Belgie	0,21	0,20	0,29
BRA – Brazílie	0,03	0,01	0,24
CRO – Chorvatsko	0,14	0,14	0,11
CZE – Česko	0,17	0,18	0,06
ESP – Španělsko	0,64	0,63	0,65

FRA – Francie	0,11	0,11	0,12
GBR – Velká Británie	0,31	0,30	0,40
GER – Německo	-0,05	-0,04	-0,04
CHI – Chile	0,12	0,11	0,12
ITA – Itálie	0,04	0,04	-6e-4
NED – Nizozemí	0,19	0,20	0,14
ROM – Rumunsko	0,14	0,15	0,10
RUS – Rusko	1,07	1,09	0,94
SER – Srbsko	0,53	0,54	0,49
SUI – Švýcarsko	0,34	0,32	0,54
SWE – Švédsko	0,36	0,37	0,25
USA – Spojené státy americké	-0,26	-0,26	-0,20

Tabulka 16 : Hodnoty regresních parametrů vysvětlující proměnné národnost na turnaji French Open

Dále jsme ukázali, že držení rakety se neprokázalo pro tenistu jako podstatný faktor, a to jak celkově, tak ani na jednotlivých turnajích.

Poslední námi zkoumanou vysvětlující proměnnou byla výška tenisty. Tato proměnná se potvrdila jako významná v celkovém hodnocení a poté i na dvou turnajích a to na French Open a US Open. Tedy na dvou protipólech námi zkoumaných turnajů, na turnajích s nejpomalejším a nejrychlejším povrchem. V tabulce (17) uvádíme hodnotu regresního parametru odpovídající statisticky významné proměnné výšku u modelů splňující podmínku nulovosti střední hodnoty náhodné složky a konstantnosti rozptylu náhodné složky.

	První dvojice výsledků	**Poslední dvojice výsledků**	**1**
	γ	γ	γ
Celkem	0,01	0,01	0,01
French Open	0,01	0,01	0,01
US Open	0,02	0,02	0,02

Tabulka 17 : Hodnoty regresních parametrů vysvětlující proměnné výška

9 Závěr

V práci bylo testováno, zda existuje rozdíl ve výkonech tenistů na jednotlivých površích. To bylo testováno pomocí párového t-testu, kde byly porovnávány výsledky tenistů vždy na dvojicích turnajů. Pokud byli do testování zahrnuti všichni tenisté, kteří se dané dvojice turnajů zúčastnili, nepodařilo se nám prokázat rozdíl mezi žádnou ze dvojic turnajů. Poté jsme, ale na příkladu Španělska ukázali, že pokud vybereme jednotlivou národnost, můžeme pozorovat rozdílnost mezi turnaji a tedy jednotlivými povrchy.

Poté jsme testovali vliv národnosti tenisty, to v jaké ruce drží raketu a jeho výšky na jeho průměrný bodový výkon za sledované období 2000-2011. A to, jak celkově, tak na jednotlivých turnajích. Pro testování závislostí tenistova výkonu na uvedených faktorech byla použita vážená metoda nejmenších čtverců. Za její pomoci jsme prokázali, že národnost tenisty se ukázala jako statisticky významná pouze na turnaji French Open. U druhého sledovaného faktoru držení rakety se nám nepodařilo prokázat žádný statisticky významný vliv. Posledním sledovaným faktorem byla výška tenisty, která se projevila statistiky významnou v celkovém hodnocení a poté na dvou dílčích turnajích, a to na French Open a US Open.

Seznam použité literatury

Reference

[1] J. Reif, *Metody matematické statistiky*, 2. upravené vydání, Plzeň, Západočeská univerzita, 2004.

[2] K. Zvára, *Regrese*, 1. vydání, Praha, Matfyzpress, 2008.

[3] B. Šedivá, *Matematické modely v ekonometrii*, ZČU, 2012, výukový text k přednáškám. Dostupné z: <http://home.zcu.cz/~sediva/mme.htm>.

[4] B. Šedivá, *Výpočtová statistika*, ZČU, 2010, výukový text k přednáškám. Dostupné z: <http://home.zcu.cz/~sediva/stav.htm>.

[5] M. Meloun, J. Militký, *Kompendium statistického zpracování dat,* 2. přepracované a rozšířené vydání, Praha, Academia, 2006.

[6] MATLAB Documentation Center [online]. [cit. 2013-5-26]. Dostupné z: <http://www.mathworks.com/help/stats/>.

[7] TenisPortal.cz [online]. [cit. 2013-5-26]. Dostupné z: <http://www.tenisportal.cz/seznam-hracu/>.

[8] Australian Open Tennis Championships 2013 - The Grand Slam of Asia/Pacific [online]. [cit. 2013-5-26]. Dostupné z: <http://www.australianopen.com/index.html>.

[9] Roland Garros - The 2013 French Open[online]. [cit. 2013-5-26]. Dostupné z: <http://www.rolandgarros.com/index.html>.

[10] 2013 Wimbledon Championships Website [online]. [cit. 2013-5-26]. Dostupné z: <http://www.wimbledon.com/en_GB/index.html>.

[11] 2013 US Open Official Site [online]. [cit. 2013-5-26]. Dostupné z: <http://www.usopen.org/>.

[12] Wikipedia. Kategorie:Grandslamové turnaje [online]. [cit. 2013-5-26]. Dostupné z: <http://cs.wikipedia.org/wiki/Kategorie:Grandslamov%C3%A9_turnaje>.

[13] Humusoft. MATLAB - Jazyk pro technické výpočty [online]. [cit. 2013-5-26]. Dostupné z: <http://www.humusoft.cz/produkty/matlab/matlab/>.

[14] Humusoft. MATLAB - Jazyk pro technické výpočty [online]. [cit. 2013-5-26]. Dostupné z: <http://www.humusoft.cz/produkty/matlab/aknihovny/statistics/>.