

Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky



Ekonometrická analýza běžeckých závodů

BAKALÁŘSKÁ PRÁCE

Studijní program: Kvantitativní metody v ekonomice

Studijní obor: Matematické metody v ekonomii

Autor: Veronika Fenclová

Vedoucí bakalářské práce: Mgr. Vladimír Holý, Ph.D.

Praha, květen 2019

Prohlášení

Prohlašuji, že jsem bakalářskou práci „Ekonometrická analýza běžeckých závodů“ vypracovala samostatně za použití v práci uvedených pramenů a literatury.

V Praze dne 6. května 2019

.....

Veronika Fenclová

Poděkování

Na tomto místě bych ráda poděkovala Mgr. Vladimíru Holému, Ph.D. za vedení a poskytnutí cenných rad k mé bakalářské práci.

Abstrakt

Bakalářská práce se zabývá ekonometrickou analýzou běžeckých závodů. Cílem této práce je najít faktory, které ovlivňují běžce při výběru běžeckých závodů a zároveň zjistit, jaký vliv tyto faktory na běžce mají. Jedná se například o vzdálenost, cenu nebo den pořádání závodu. Všechny faktory jsou blíže specifikovány v první kapitole. Data jsou sesbírána za rok 2018. V další části práce je stručně popsána teorie k modelům, kde závisle proměnná nabývá nezáporných celočíselných hodnot. Jako první je uvedena Poissonova regrese a s tím související možný výskyt nadměrné disperze. Jako řešení tohoto problému je v této práci uvažována negativně binomická regrese, známá jako NB2. V poslední části jsou sestaveny modely. První model je odhadnut Poissonovou regresí. Z důvodu výskytu nadměrné disperze je další model odhadnut za pomoci negativně binomické regrese. Do třetího modelu byla přidána skupina proměnných nazvaná reklama. Na základě posledního modelu, který obsahuje pouze statisticky významné proměnné, jsou otestovány hypotézy a interpretovány výsledky. Podle sestaveného modelu měla většina faktorů očekávaný vliv.

Klíčová slova

běžecké závody, čítací modely, ekonometrická analýza

JEL klasifikace

C 250

Abstract

Bachelor thesis deals with econometric analysis of running races. The aim of the thesis is to find factors influencing runner's choice of the races and to find out how important these factors are. Factors are for example course length, price or date of the race. These factors are described in the first section. Data have been collected for 2018. The theory of count data is briefly described in the next part; the first is Poisson regression and possible problem of overdispersion. This thesis considers negative binomial regression (NB2) as a solution of this problem. The last section contains models. For the first model estimation the Poisson regression is used. This model shows significant overdispersion; to handle the overdispersion the NB2 is used for the next model estimation. Variables called advertisement were added to the third model. The last model contains only statistically significant predictors. Then the hypotheses are tested, and results are interpreted. Most of the significant variables had the expected effect.

Keywords

count data, econometric analysis, running races

JEL Classification

C 250

Obsah

Úvod	7
1 Data	8
1.1 Původ dat	8
1.2 Závisle proměnná	8
1.3 Nezávisle proměnné	9
2 Metodologie	13
2.1 Poissonova regrese	13
2.2 Nadměrná disperze	14
2.3 Negativně binomická regrese	14
3 Empirická analýza	17
3.1 Modely	17
3.2 Interpretace výsledků	24
3.3 Testování hypotéz	26
3.4 Diskuze	27
Závěr	29
Použitá literatura	30

Úvod

V posledních letech se klade velký důraz na zdravý životní styl. Možná i tento fakt přispěl k tomu, že běhání se stává čím dál více populárnější. Dalším důvodem může být menší finanční zátěž, než je tomu u jiných sportů. Tato práce se zabývá účastí na běžeckých závodech. Proč některé závody přilákají na start tisíce běžců a u jiných může počet běžců obsadit medailové pozice? Ovlivňuje rozhodování pouze fakt kde, kdy se daný závod koná a jeho vzdálenost, nebo i ve sportu má velký vliv reklama? Skupina proměnných, do které patří například vzdálenost, ročník nebo místo konání závodu, je v této práci označena za parametry závodů. Druhá skupina faktorů, která je uvažována, se jmenuje reklama. Hlavním cílem této práce je sestavit takový ekonometrický model, za pomoci kterého se podaří zodpovědět nejenom výše zmíněné otázky, ale i níže uvedené hypotézy.

H1: Závodníci si vybírají závody pouze na základě parametrů závodů.

H2: Konání přidružených akcí přiláká na hlavní závod více běžců.

H3: Vysoké startovné negativně ovlivňuje počet účastníků.

H4: Běžci dávají přednost závodům v krajských městech než v jiných lokalitách.

1 Data

1.1 Původ dat

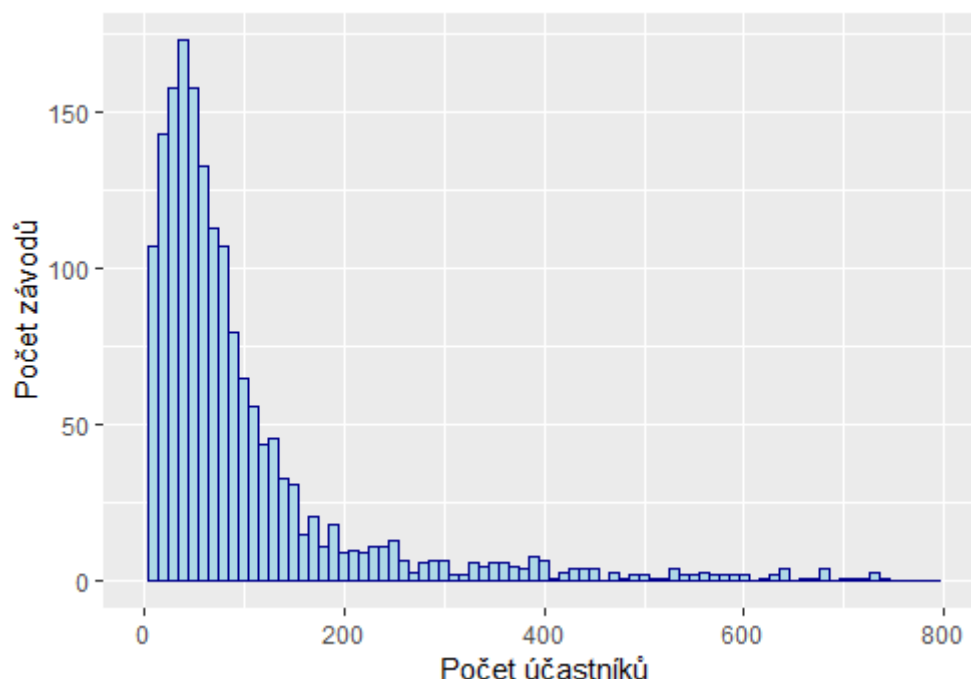
Hlavním zdrojem dat, která jsou použita v této práci, je běžecká termínovka [1]. Z tohoto zdroje byla vzata data závodů, vzdálenosti a místa konání. Zbytek dat byl doplněn z oficiálních stránek jednotlivých závodů, portálů uvádějící oficiální výsledky nebo termínovek v podobě časopisů. Data obsahují vybrané běžecké závody konané na území České republiky za rok 2018. Jedná se o závody jednotlivců, kterých se mohou účastnit jak muži, tak ženy. Soubor se skládá z 1779 pozorování.

Tabulka 1: Základní statistické údaje o kvantitativních proměnných

Proměnná	Minimum	1.kvartil	Medián	Průměr	3.kvartil	Maximum
Účastníci	1,0	37,0	66,0	155,1	121,5	10 118,0
Vzdálenost	1,0	5,0	9,0	12,2	12,0	161,0
Ročník	1,0	3,0	6,0	12,3	16,0	122,0
Cena	0,0	50,0	100,0	183,3	250,0	2 350,0
Google	1,0	5,0	9,0	66,1	72,0	2060,0

1.2 Závisle proměnná

Účastníci – jedná se o kvantitativní proměnnou, která udává počet běžců, kteří doběhli do cíle. Tato proměnná nezahrnuje běžce, kteří neodstartovali nebo byli diskvalifikováni. Z tabulky (Tabulka1), je patrné, že tato proměnná obsahuje extrémní hodnoty, jelikož průměrný počet běžců je 155, zatímco medián je 66 běžců. I samotná maximální hodnota, která čítá 10 118 běžců, poukazuje na velké rozdíly. Tato extrémní hodnota udává účast na pražském půlmaratonu, kterého se účastní i mezinárodní běžci, kdežto ostatních závodů především Češi. Pražský půlmaraton není jediným závodem s vysokou účastí. Je zde několik dalších závodů, které mají několikanásobně vyšší účast, než je průměr. Jedná se především o závody pořádané nejznámějšími organizátory v České republice, jakými jsou RunCzech nebo RunTour. Nesmí se opomenout ani některé závody pořádané pro charitu, kterých se také účastní mnoho běžců. U většiny závodů je stanovena kapacita běžců, kteří se mohou daného závodu zúčastnit, avšak není obvyklé, že by kapacita byla naplněna. Rozdělení četností této proměnné je vyobrazeno na obrázku (Obrázek1). Pro lepší vizualizaci byly vynechány závody s účastí nad 800 běžců. Jedná se o 43 závodů, které nejsou v tomto histogramu zachyceny.



Obrázek 1: Histogram počtu účastníků

1.3 Nezávisle proměnné

Existuje mnoho faktorů, které ovlivňují účast na běžeckých závodech. Pro analýzu v této práci bylo zvoleno následujících 10 vysvětlujících proměnných, které se dají rozdělit do dvou skupin. První skupinu tvoří parametry závodů. Do této skupiny spadají všechny níže uvedené proměnné, kromě proměnných ocenění a Google, které tvoří druhou skupinu nazvanou reklama. Parametry závodů je skupina proměnných, kterými se daný závod dá popsat neboli můžeme odpovědět například na základní otázky typu kde, kdy se daný závod koná a jaká vzdálenost musí být překonána. Reklama je skupina, kde velkou roli hrají sponzoři a prestiž jednotlivých závodů. Základní statistické údaje ke kvantitativním proměnným jsou uvedeny v tabulce (Tabulka1) a základní informace k dummy proměnným jsou k vidění v tabulce (Tabulce2).

Vzdálenost – neboli délka trati závodů, tato proměnná je uvedena v kilometrech. Nejkratší napozorovanou vzdáleností byl závod na 1 km. Takto krátkých závodů se účastní především sprinteři. Naopak nejdelší závod byl dlouhý celých 161 km, jedná se o ultramaraton s vypovídajícím názvem TransMoravský Masochistický Terénní Běh. Většina běžců běhá vzdálenosti do 20 km, jelikož delší závody jsou náročnější na přípravu a následnou regeneraci po závodě. Tomuto odpovídá i fakt, že průměrná vzdálenost závodů je 12 km. Zbylé popisné charakteristiky jsou uvedeny v tabulce (Tabulka1). U této proměnné bude očekáván nejspíše negativní vliv na účast, jelikož zaběhnout maraton není tak jednoduché, jako uběhnout kratší vzdálenosti.

Ročník – udává, pokudikáté se tento závod pořádá. Nejstarším závodem, který byl v datech zaznamenán, je 10 km závod Běchovice-Praha, pořádaný Českou spořitelnou. Během roku 2018 přibýlo 152 nových závodů. Proměnná ročník bude mít nejspíše pozitivní vliv na účast,

protože s každým dalším rokem se o závodě může dozvědět více lidí, a to i za pomoci pozitivních reakcí od běžců.

Cena – neboli startovné, je uvedena v Kč. Běžecké závody nemají obvykle jednotnou cenu, často záleží na datu přihlášení. Většina běžců se přihlašuje předem, proto tato proměnná obsahuje především startovné předem. U některých závodů je možné platit startovné pouze na místě, zde pak byla brána tato částka. Další skupinu tvořily běhy, kde startovné bylo dobrovolné. U takových závodů byla zadána částka 0. Dobrovolné startovné je většinou na malých závodech pořádaných běžeckými spolky, naopak nejdražší startovné je na Volkswagen Maraton Praha.



Obrázek 2: Korelace mezi kvantitativními proměnnými

Na obrázku (Obrázek2) je zachycena korelace mezi kvantitativními proměnnými, které jsou uvedeny na diagonále. V levé části je uvedena hodnota korelačního koeficientu a v pravé části grafické znázornění této závislosti. U přímé závislosti je vyobrazen modrý kruh, u nepřímé se jedná o červený. Síla závislosti je pak vyobrazena velikostí a odstínem dané barvy. Největší přímá korelace mezi kvantitativními proměnnými je tedy mezi cenou a účastníky. Naopak největší nepřímá závislost je mezi cenou a ročníkem. Mezi ročníkem a vzdáleností není žádný lineární vztah. Pokud se v datech vyskytují odlehlá pozorování, mohou ovlivnit korelační koeficient.

Tabulka 2: Průměrné hodnoty závisle proměnné v závislosti na dummy proměnné

Proměnná	Průměr		
	Celkem	0	1
Pohár	0,34	119,36	225,72
Dětský závod	0,55	179,38	135,36
Den	0,78	92,83	172,02
Krajské město	0,30	102,87	275,75
Dopoledne	0,57	174,65	140,21
Odpoledne	0,38	160,06	146,91
Večer	0,05	142,90	383,27
Ocenění	0,09	116,12	554,49

Pohár – jedná se o dummy proměnnou, kde hodnota 1 udává, že závod je součástí poháru. Znamená to, že během roku se na různých místech běží závody pořádané jedním organizátorem. V opačném případě nabývá hodnoty 0. Příklady poháru jsou závody od organizátorů závodů Běžíme na Hrad, Winter Run a Běhej lesy.

Dětský závod – tato proměnná nabývá hodnot 0 a 1. Pokud součástí hlavního závodu je i jiný závod, nabývá tato proměnná hodnoty 1. Jiným závodem je zde myšlen například dětský závod, který tvoří největší skupinu, po které nese tato proměnná název. Dalšími doprovodnými závody jsou například závody na koloběžkách, nordic walking nebo canicross.

Den – další proměnná ze skupiny dummy proměnných, kde hodnota 1 značí pořádání závodu o víkendu nebo o státním svátku. Pokud je běžecký závod konán v týdnu, nabývá tato proměnná hodnoty 0. O víkendu nebo státním svátku je pořádáno 78 % závodů, zároveň o víkendu je průměrná návštěvnost o 80 běžců větší než průměrná návštěvnost v týdnu (Tabulka2).

Krajské město – hodnota 1 je přiřazena závodem konaným v krajském městě v opačném případě je přiřazena hodnota 0. V krajském městě bylo pořádáno 30 % závodů (Tabulka2). Jsou zde konány především velké závody od již zmiňovaných organizátorů, jakými jsou RunCzech a RunTour. V krajských městech jsou také často pořádány charitativní běhy.

Čas – jedná se o kategoriální proměnnou, skládající se ze 3 kategorií: dopoledne (6:00-11:59), odpoledne (12:00-18:59) a večer (19:00-0:00). V čase mezi 0:01-5:59 nebyl odstartován žádný závod, proto tento čas nebyl uvažován. V 6 hodin se vybíhá převážně na delší tratě. Jedná se především o maratony a o půlnoci byl odstartován jediný závod a to Silvestrovský. Nadpoloviční většina závodů (Tabulka2) startuje v dopoledních hodinách.

Ocenění – tato dummy proměnná udává, zda byl závod, respektive jeho pořadatel, oceněn Českým atletickým svazem za nejlepší závod roku 2017. Pokud byl oceněn pořadatel, tak u všech závodů, jím pořádaných, nabývá ocenění hodnoty 1.

Google – udává, kolik stránek je nalezeno pomocí vyhledávače Google po zadání závodu ve tvaru „Název závodu“ místo a datum (např. „Běh pro Paměť národa“ Třebíč 9. 6. 2018). Tento způsob byl zvolen z důvodu, že některé závody nesou stejný název například závody, které jsou součástí poháru, nebo název Mikulášský běh neslo také více závodů konaných v různých dnech a na různých místech. Tato proměnná by měla zachytit vliv reklamy některých závodů. Čím více stránek je nalezeno, tím je pro běžce snadnější se o závod dozvědět.

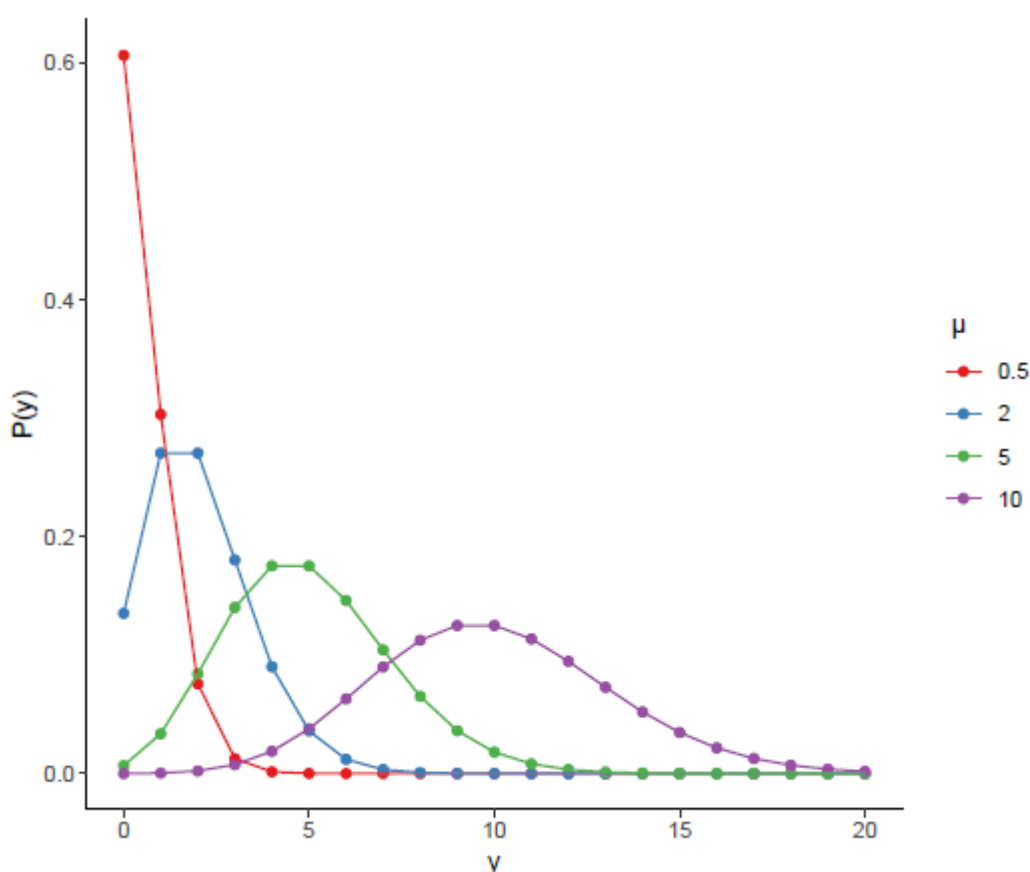
2 Metodologie

V této sekci budou přiblíženy postupy, kterými byly v kapitole Empirická analýza sestaveny a následně odhadnuty ekonometrické modely. Jedná se o modely, kde závisle proměnná nabývá pouze nezáporných celočíselných hodnot. V této práci jsou využity tzv. čítací modely (count data), kam spadá Poissonova a negativně binomická regrese. Data bývají často kladně sešikmena a zatížena heteroskedasticitou [6].

2.1 Poissonova regrese

Pravděpodobnostní funkce Poissonova rozdělení má tvar

$$P(Y = y) = \frac{e^{-\mu} \cdot \mu^y}{y!}, \quad y = 0, 1, \dots \quad (1)$$



Obrázek 3: Poissonovo rozdělení

Jde o rozdělení s jedním parametrem, a to střední hodnotou a rozptylem μ [2,6]. Na obrázku (Obrázek3) je zachyceno Poissonova rozdělení pro různé parametry μ .

Základním předpokladem tohoto modelu je rovnost podmíněné střední hodnoty a rozptylu.

$$\begin{aligned} V[y_i | \mathbf{x}_i] &= \mu_i, \\ E[y_i | \mathbf{x}_i] &= \mu_i. \end{aligned} \quad (2)$$

Pro logaritmickou verzi je podmíněná střední hodnota definována

$$\log(\mu) = \mathbf{x}\beta. \quad (3)$$

Druhým předpokladem jsou nezávislá pozorování. Tento předpoklad vychází z toho, že k událostem dochází nezávisle na čase [3]. Může existovat i dynamická závislost mezi událostmi neboli předchází výskyt události, například nemoci, může zvýšit pravděpodobnost výskytu stejných nebo podobných událostí [2]. Poissonův model je odhadován metodou maximální věrohodnosti. Tato metoda poskytuje konzistentní odhady za předpokladu správné specifikace podmíněného průměru, zároveň není vyžadováno, aby závisle proměnná měla Poissonovo rozdělení. Ke správnému určení směrodatných chyb a následné statistické významnosti proměnných je zapotřebí nejen správné specifikace podmíněného průměru, ale také podmíněného rozptylu [3].

2.2 Nadměrná disperze

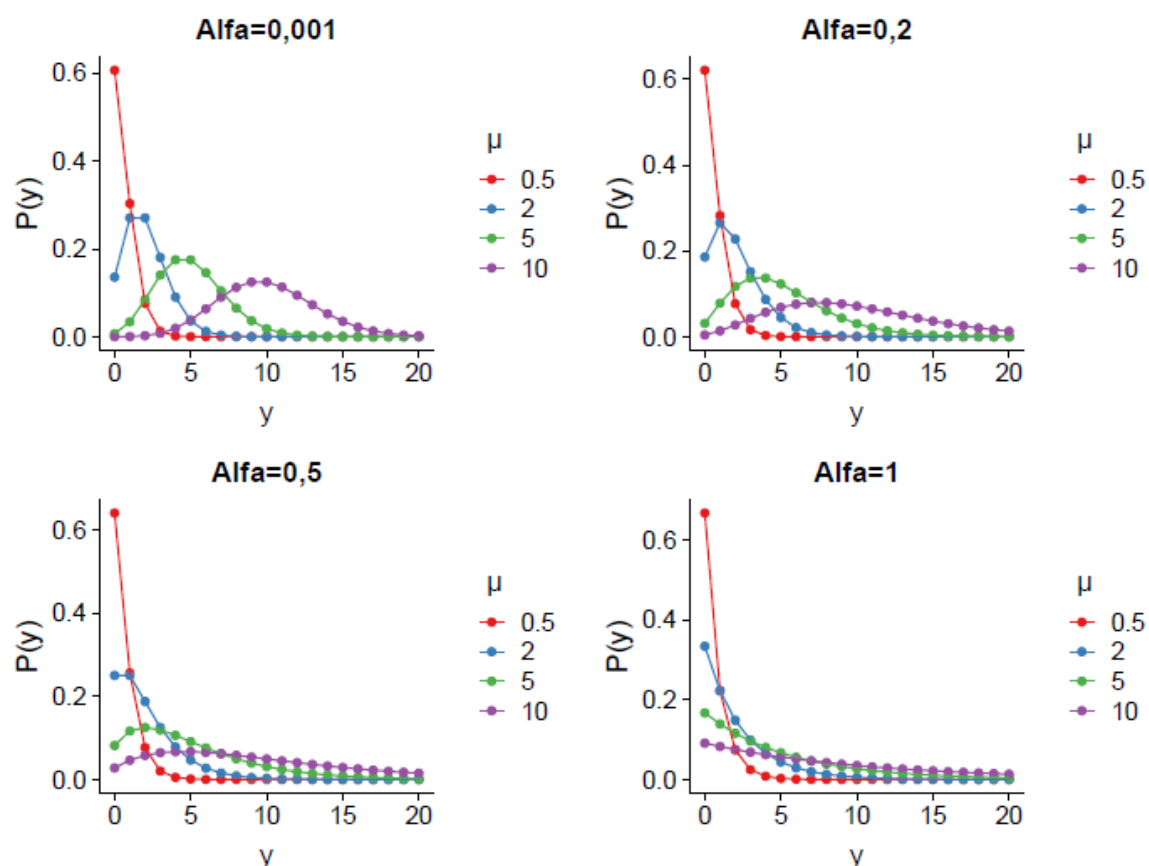
V reálných datech je často porušena podmínka rovnosti střední hodnoty a rozptylu (2). V případě, že je rozptyl větší jak střední hodnota, mluvíme o nadměrné disperzi. Tento jev může být způsoben například pozitivní korelací mezi výsledky, což značí porušení předpokladu [6]. Nadměrná disperze neovlivní odhady parametrů β , avšak odhady směrodatných chyb jsou podhodnocené, jak již bylo zmíněno v předchozím odstavci. Někdy se hovoří o takzvané zdánlivé nadměrné disperzi, kterou mohou způsobit například odlehlá pozorování nebo nezahrnutí důležité vysvětlující proměnné. V případě, že se v datech vyskytuje tento problém, lze to řešit několika způsoby. Pokud se jedná o zdánlivou nadměrnou disperzi, je zde možnost přidání nezahrnuté vysvětlující proměnné, přidání interakcí nebo transformace proměnných [6]. Pokud i po těchto krocích se vyskytuje nadměrná disperze, je možné použít jiné modely. Testování předpokladu (2) je podrobně popsáno v literatuře [4].

2.3 Negativně binomická regrese

Pravděpodobnostní funkce tohoto rozdělení je dána

$$P(Y = y) = \binom{r+y-1}{r-1} (1-p)^r p^y, \quad y = 0, 1, 2, \dots, \quad (4)$$

kde $0 < p < 1$ a $r > 0$, zároveň $r = \frac{1}{\alpha}$ a $p = \frac{\alpha\mu}{(1+\alpha\mu)}$ [3].



Obrázek 4: Negativně binomické rozdělení

Na obrázku (Obrázek4) jsou uvedeny negativně binomická rozdělení pro různé hodnoty parametru α . Jednou z možností, jak se vypořádat s nadměrnou disperzí, je použití negativně binomické regrese. Existuje několik různých druhů této regrese. Záleží na definování vztahu mezi střední hodnotou a rozptylem. V této práci bude využita standardní negativně binomická regrese označovaná jako NB2. V této formě je střední hodnota a rozptyl ve vztahu

$$\begin{aligned}
 E[Y_i | x_i] &= \mu_i \\
 \text{Var}[Y_i | x_i] &= \mu_i \left(1 + \frac{\mu_i}{\theta} \right) = \mu_i (1 + \alpha \mu_i)
 \end{aligned} \tag{5}$$

Poissonova regrese může být chápána jako speciální případ negativně binomické regrese. Pokud se parametr α blíží k 0, pak je negativně binomický model zaměnitelný za Poissonův model, jak již bylo patrné z grafů (Obrázek3, Obrázek4). Další možné parametrizace rozptylu negativně binomického modelu jsou uvedeny [6].

Ke zhodnocení modelu mohou pomoci rezidua, která mohou upozornit na odlehlá a vlivná pozorování. U modelů se závisle proměnnou, která nabývá pouze nezáporných celočíselných hodnot je využíváno několik druhů reziduí. Mezi nejznámější rezidua, využívaných v těchto modelech, patří například Pearsonovy nebo devianční rezidua. Reziduální analýza může pomoci identifikovat odlehlá pozorování, nebo špatnou specifikaci modelu. Dalšími

využívanými rezidui jsou Studentova rezidua, která jsou používána k identifikaci vlivných pozorování společně s Cookovou vzdáleností [6].

K ověření správné specifikace modelu se využívá několika metod. Nedá se jednoznačně určit, který z nástrojů je nejvhodnější. Jedním z nejznámějších nástrojů je koeficient determinace R^2 , který je využíván především v metodě nejmenších čtverců. V nelineárních modelech se využívá pseudo- R^2 . Nízké hodnoty pseudo- R^2 mohou naznačovat, že model není správně specifikován. Existuje několik možností výpočtu tohoto koeficientu [3]. Dalším nástrojem sloužícím k porovnání modelů je například Pearsonova statistika, ta se používá především u Poissonova modelu k odhalení nadměrné disperze. Další možnosti hodnocení modelu je devianční statistika, test věrohodnostního poměru, nebo χ^2 test dobré shody [3,6].

Závisle proměnná je uvedena v logaritmické podobě. Pro přesnější odhady koeficientů je vhodné je přepočítat pomocí exponenciály.

$$\hat{\beta}_j^* = e^{\hat{\beta}_j} - 1. \quad (6)$$

Dále se mohou interpretovat mezní efekty, které udávají pravděpodobnostní změnu výsledku s jednotkovou změnou vysvětlující proměnné za předpokladu, že zbylé proměnné jsou neměnné [6]. V této práci jsou uváděny průměrné mezní efekty, které se vypočítají jako

$$AME = \hat{\beta}_k \bar{y}. \quad (7)$$

Tato hodnota udává o kolik v průměru vzroste nebo se sníží závisle proměnná při jednotkové změně vysvětlující proměnné [6].

3 Empirická analýza

3.1 Modely

Modely, které jsou uvedeny v této části, byly odhadnuty v programu R. K odhadům byla využita funkce `glm` a `glm.nb`. Ve funkci `glm.nb` je místo parametru α uváděn parametr θ . Tyto dva parametry jsou ve vztahu

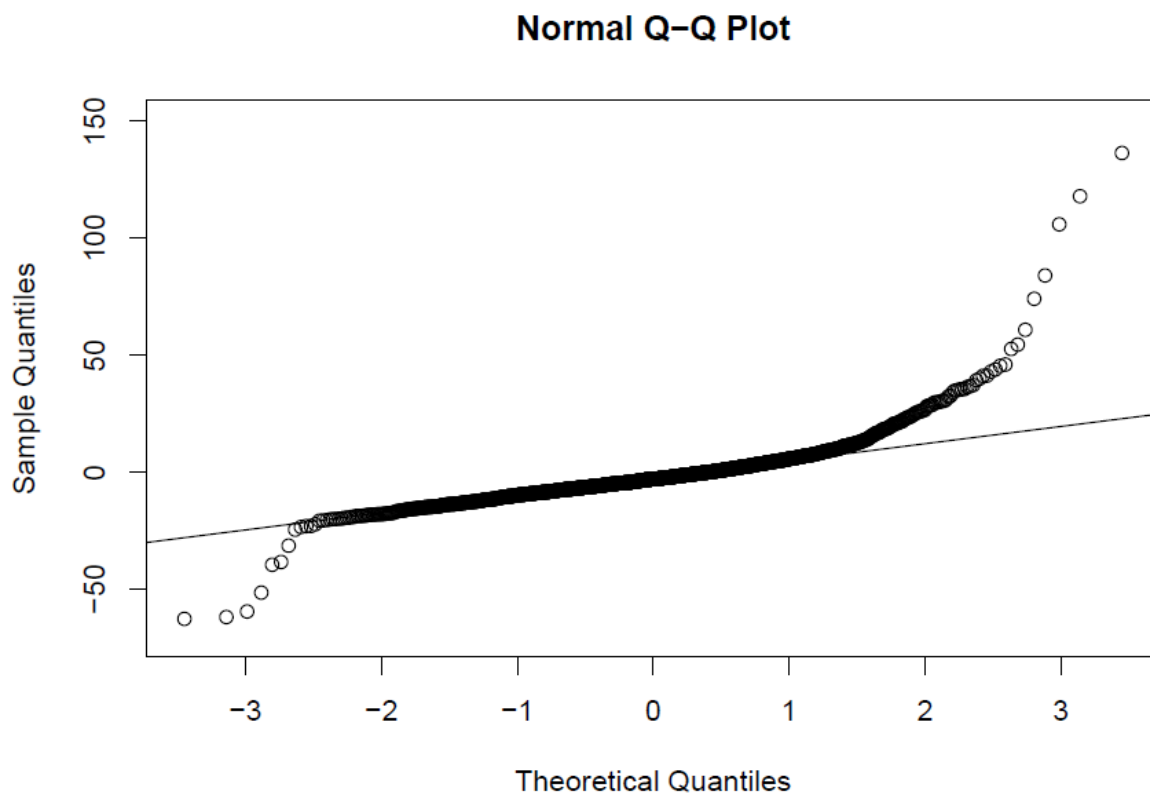
$$\alpha = \frac{1}{\theta}. \quad (8)$$

Přestože je zvolen jiný parametr, nejedná se o jiný model z důvodu výskytu parametru α v rozptylu nikoli ve střední hodnotě.

Model 1 obsahuje proměnné nazvané parametry závodů a je odhadnut Poissonovou regreseí. Výstup z tohoto modelu je uveden v tabulce (Tabulka 3). U kategoriální proměnné čas byla zvolena referenční kategorie večer. Prvně musí být ověřen základní předpoklad této regrese, a to rovnost podmíněné střední hodnoty a rozptylu. Testuje se nulová hypotéza o nulovosti koeficientu α z rovnice (5). Na 5 % hladině významnosti můžeme zamítnout tuto nulovou hypotézu (p -hodnota = $1,47 \times 10^{-8}$). Data jsou tedy zatížena nadměrnou disperzí, proto směrodatné chyby, stejně jako p -hodnoty uvedené v tomto modelu, jsou nesprávné. I rozpětí reziduí (Obrázek 5) napovídá o tom, že takto sestavený model, odhadnutý pomocí Poissonovy regrese, není správný. Pro tato data je vhodnější použití negativně binomické regrese.

Tabulka 3: Model 1 - Poissonova regrese

Proměnná	Odhad	Směrod. chyba	P-hodnota
Konstanta	4,565	$7,95 \times 10^{-3}$	$2,00 \times 10^{-16}$
Vzdálenost	-0,024	$1,98 \times 10^{-4}$	$2,00 \times 10^{-16}$
Pohár	0,336	$4,14 \times 10^{-3}$	$2,00 \times 10^{-16}$
Dětské závody	-0,020	$4,05 \times 10^{-3}$	$2,00 \times 10^{-16}$
Ročník	0,010	$1,21 \times 10^{-4}$	$2,00 \times 10^{-16}$
Den	0,877	$6,43 \times 10^{-3}$	$2,00 \times 10^{-16}$
Dopoledne	-1,246	$6,64 \times 10^{-3}$	$2,00 \times 10^{-16}$
Odpoledne	-0,895	$6,43 \times 10^{-3}$	$2,00 \times 10^{-16}$
Cena	0,002	$4,28 \times 10^{-6}$	$2,00 \times 10^{-16}$
Krajské město	0,715	$4,06 \times 10^{-3}$	$2,00 \times 10^{-16}$
Nulová deviance	597 063	Stupně volnosti	1778
Reziduální deviance	247 205	Stupně volnosti	1769

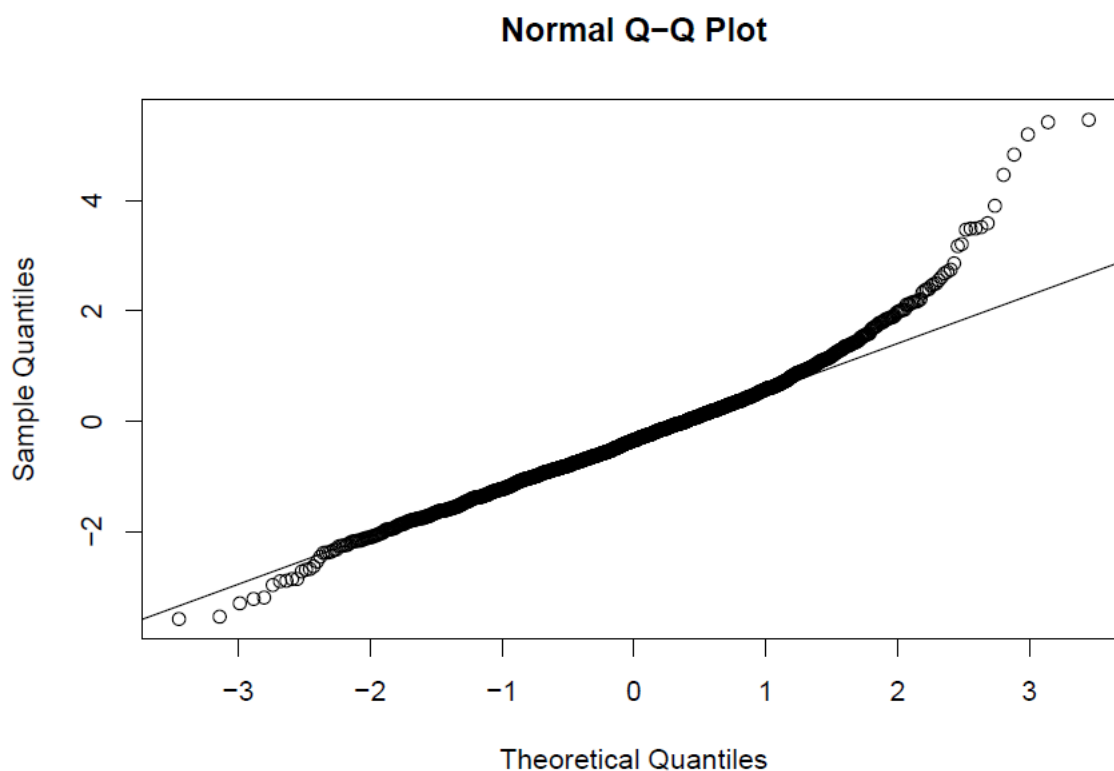


Obrázek 5: Analýza reziduí modelu 1 (devianční rezidua)

Tabulka 4: Model 2 - NB2, parametry závodů

Proměnná	Odhad	Směrod. chyba	P-hodnota
Konstanta	4,176	0,1347	$2,00 \times 10^{-16}$
Vzdálenost	-0,017	0,0022	$5,77 \times 10^{-15}$
Pohár	0,207	0,0563	$2,44 \times 10^{-4}$
Dětské závody	-0,007	0,0539	$8,88 \times 10^{-1}$
Ročník	0,010	0,0018	$1,54 \times 10^{-8}$
Den	0,314	0,0795	$7,86 \times 10^{-5}$
Dopoledne	-0,567	0,1331	$2,27 \times 10^{-5}$
Odpoledne	-0,401	0,1269	$1,16 \times 10^{-3}$
Cena	0,003	0,0001	$2,00 \times 10^{-16}$
Krajské město	0,335	0,0059	$4,14 \times 10^{-8}$
Nulová deviance	4054,9	Stupně volnosti	1778
Reziduální deviance	1963,1	Stupně volnosti	1769

Model 2 je odhadnut pomocí negativně binomické regrese. Tento model obsahuje, stejně jako model 1, proměnné nazvané parametry závodů. Takto odhadnutý model je uveden v tabulce (Tabulka4). Proměnná dětské závody se zde jeví jako statisticky nevýznamná. Před testováním významnosti jednotlivých proměnných je vhodné posoudit model jako celek. Interpretace reziduí není tak jednoznačná jako u jiných modelů. Z grafu (Obrázek6) je vidět, že rezidua nejsou normálně rozdělena. Pro negativně binomickou regresi, nemusí být splněn předpoklad o normálním rozdělení náhodné složky, ale může to naznačovat výskyt odlehlých pozorování. Rezidua v pravé horní části naznačují výskyt odlehlých pozorování, zároveň i v levé dolní části se nacházejí rezidua pod přímkou, zde však nedochází k tak velkému vychýlení. Přítomnost extrémních hodnot byla v datovém souboru zřejmá již z hodnot v tabulce (Tabulka1). Model neobsahuje proměnnou, která by vysvětlila tyto značné rozdíly mezi závody. Z tohoto důvodu se model 2 jeví jako nesprávně specifikovaný. Hypotéza *H1*, o tom že by se podařila vysvětlit závisle proměnná účast, pouze na základě parametrů závodů, je zamítnuta. Nezachyceným faktorem bude nejspíše vliv reklamy a sponzorů.



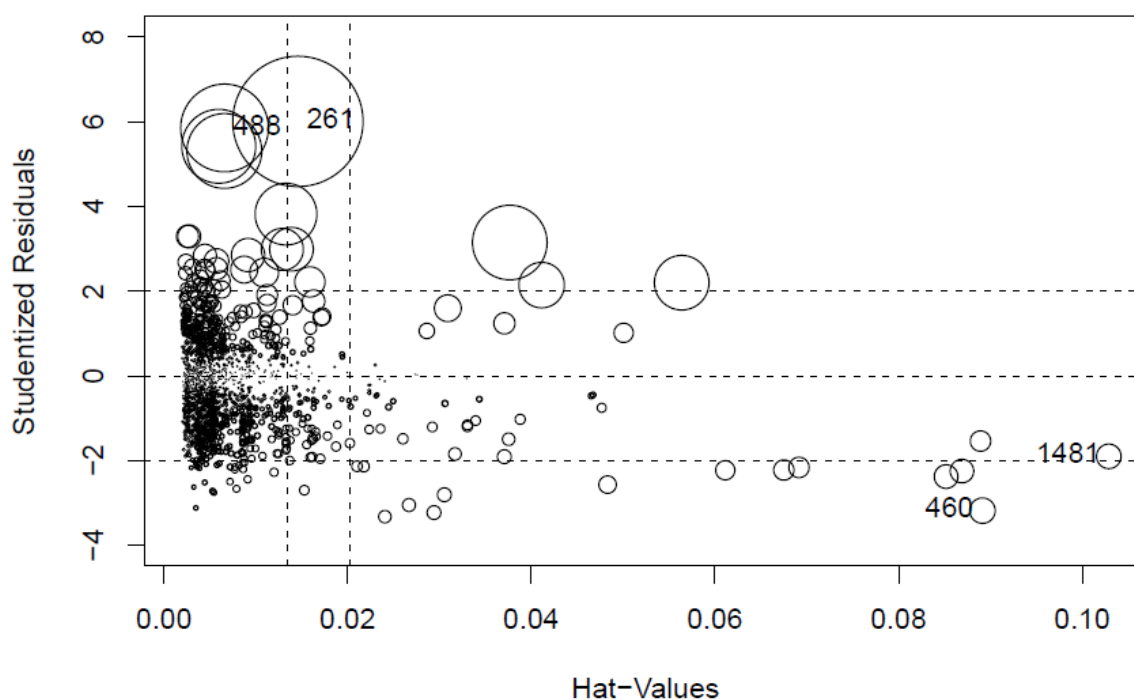
Obrázek 6: Analýza reziduí modelu 2 (devianční rezidua)

Model 3 je rozšířen o proměnné ze skupiny reklama, kam spadá proměnná Google a ocenění. Ocenění je tzv. dummy proměnná a udává, zda byl závod oceněn Českým atletickým svazem, jak již bylo zmíněno v sekci Data. Většina z oceněných závodů spadá mezi ty nejnavštěvovanější. Proměnná Google udává počet vyhledaných stránek pomocí vyhledávače Google. Očekává se, že závody s více odkazy budou i více navštěvované. Z tabulky (Tabulka5) je zřejmé, že obě přidané proměnné se jeví na 0,1 % hladině významnosti jako statisticky významné. Proměnná ocenění značí, že pokud byl závod oceněn, v průměru přijde o necelých 61 % více lidí, než kdyby závod oceněn nebyl za

předpokladu, že ostatní proměnné se nemění. Proměnná Google má také pozitivní vliv na účast. S každou další stránkou se zvýší počet účastníků v průměru o 0,2 %. Proměnná dětské závody se opět jeví jako statisticky nevýznamná.

Tabulka 5: Model 3 - NB2, proměnné reklama

Proměnná	Odhad	Směrod. chyba	P-hodnota
Konstanta	4,024	0,1285	$2,00 \times 10^{-16}$
Vzdálenost	-0,016	0,0021	$1,78 \times 10^{-14}$
Pohár	0,214	0,0533	$6,20 \times 10^{-4}$
Dětské závody	-0,039	0,0511	$4,49 \times 10^{-1}$
Ročník	0,008	0,0017	$2,79 \times 10^{-6}$
Den	0,363	0,0773	$2,84 \times 10^{-6}$
Dopoledne	-0,432	0,1261	$3,29 \times 10^{-4}$
Odpoledne	-0,373	0,1205	$2,00 \times 10^{-3}$
Cena	0,003	0,0001	$2,00 \times 10^{-16}$
Krajské město	0,257	0,0561	$4,77 \times 10^{-6}$
Ocenění	0,472	0,0966	$1,09 \times 10^{-6}$
Google	0,002	0,0001	$2,00 \times 10^{-16}$
Nulová deviance	4473,9	Stupně volnosti	1778
Reziduální deviance	1945,7	Stupně volnosti	1767



Obrázek 7: Odlehlá pozorování modelu 3

Tabulka 6: Porovnání modelu 3 a modelu 4 bez odlehlých pozorování

Proměnná	Model 3 Odhad	Model 4 Odhad
Konstanta	4,024*** (0,1285)	3,854*** (0,0977)
Vzdálenost	-0,016*** (0,0021)	-0,014*** (0,0015)
Pohár	0,214*** (0,0533)	0,195*** (0,0402)
Dětské závody	-0,039 (0,0511)	-0,009 (0,0386)
Ročník	0,008*** (0,0017)	0,008*** (0,0013)
Den	0,363*** (0,0773)	0,391*** (0,0584)
Dopoledne	-0,432*** (0,1261)	-0,334*** (0,0956)
Odpoledne	-0,373** (0,1205)	-0,231* (0,0913)
Cena	0,003*** (0,0001)	0,003*** (0,0001)
Krajské město	0,257*** (0,0561)	0,223*** (0,0424)
Ocenění	0,472*** (0,0966)	0,530*** (0,0729)
Google	0,002*** (0,0001)	0,002*** (0,0001)
N	1779	1775

Poznámka: (i) *p<0,05, **p<0,01, ***p<0,001

Model 3 obsahuje odlehlá pozorování (Obrázek7), která se nacházejí mimo vyhraněné meze. Ze všech pozorování byla 4 pozorování vyhodnocena jako nejodlehlejší a zároveň nejvlivnější. Největší hodnotu Cookovy vzdálenosti (0,4) a zároveň největší hodnoty Studentových reziduí (6,01) nabývá závod Noční běh pro Světlušku pořádaný v Praze. Tento

závod patří mezi závody s vysokou účastí, a zároveň nespadá do kategorie oceněných závodů. Zbylá 3 pozorování nepřesahují hodnotu 0,2 Cookovy vzdálenosti, avšak hodnoty Studentových reziduí jsou vysoké.

Z odhadů regresních koeficientů (Tabulka 6) je zřejmé, že model 4 bez 4 odlehlých pozorování se výrazně neliší od modelu 3. Jediný výrazný rozdíl je u proměnné dětské závody, ale tato proměnná se jeví v těchto dvou modelech jako statisticky nevýznamná. Druhá proměnná, u které se liší odhady regresních koeficientů, je kategorie odpoledne u proměnné čas. V modelu 3 je tato kategorie statisticky významná na 1 % hladině významnosti, zatímco v modelu 4 na 5 % hladině významnosti. Již Cookova vzdálenost naznačovala, že tato pozorování nemají velký vliv na odhady parametrů. Interpretace těchto 2 modelů je podobná. Z těchto důvodů se bude nadále vycházet z modelu 3.

V modelu 3 se proměnná dětský závod jeví jako statisticky nevýznamná. Na základě testu věrohodnostního poměru nezamítáme nulovou hypotézu o statistické nevýznamnosti této proměnné (Tabulka 7).

Tabulka 7: Test věrohodnostního poměru

Model	Théta	Stupně volnosti	2xlog-likelihood	Test	P-hodnota
Model 3	1,597	1768	-19 729,56	-	-
Model 4	1,598	1767	-19 728,59	1 vs 2	0,33

Model 5 (Tabulka 8) obsahuje pouze statisticky významné proměnné. Jsou zde uvedeny nejen odhady regresních koeficientů, ale také průměrné mezní efekty. V tomto modelu je parametr $\alpha = 0,63$. Odhadnutý model má následující tvar:

$$\log(\hat{y}_i) = 4,01 - 0,016 \cdot x_1 + 0,214 \cdot x_2 + 0,008 \cdot x_3 + 0,354 \cdot x_4 - 0,455x_5 - \\ - 0,379 \cdot x_6 + 0,003 \cdot x_7 + 0,259 \cdot x_8 + 0,467 \cdot x_9 + 0,002 \cdot x_{10}$$

Koeficienty v modelu 5 (Tabulka 8) jsou uváděny v logaritmické formě. Pro přesnější a přehlednější výsledky byly koeficienty z modelu 3 přepočítány pomocí exponenciály a spolu s nimi jsou uvedeny i 95 % intervaly spolehlivosti (Tabulka 9). Nejužší intervaly jsou u proměnných Google a cena naopak proměnná ocenění má velice široký interval spolehlivosti. Očekávaný nárůst účastníků může být od 38 % do 84 % běžců.

Tabulka 8: Model 5 - odhady regresních koeficientů a průměrné mezní efekty

Model 5						
Proměnná	Regresní koeficienty			Průměrné mezní efekty		
	Odhad	Směrod. chyba	P-hodnota	Odhad	Směrod. chyba	P-hodnota
Konstanta	4,009	0,1274	2,00x10 ⁻¹⁶	-	-	-
Vzdálenost	-0,016	0,0021	2,15x10 ⁻¹⁴	-9,08	1,70	9,82x10 ⁻⁸
Pohár	0,214	0,0536	6,40x10 ⁻⁵	113,42	28,31	6,16x10 ⁻⁵
Ročník	0,008	0,0017	2,96x10 ⁻⁵	4,56	0,99	3,86x10 ⁻⁶
Den	0,354	0,0771	4,58x10 ⁻⁶	170,65	35,06	1,13x10 ⁻⁶
Dopoledne	-0,455	0,1267	3,40x10 ⁻⁴	-307,85	89,36	5,71x10 ⁻⁴
Odpoledne	-0,379	0,1209	1,73x10 ⁻³	-187,22	48,57	1,16x10 ⁻⁴
Cena	0,003	0,0001	2,00x10 ⁻¹⁶	1,52	0,22	1,96x10 ⁻¹¹
Krajské město	0,259	0,0561	4,38x10 ⁻⁶	135,33	30,70	1,04x10 ⁻⁵
Ocenění	0,467	0,0968	1,56x10 ⁻⁶	237,77	52,31	5,48x10 ⁻⁶
Google	0,002	0,0001	2,00x10 ⁻¹⁶	1,02	0,15	2,50x10 ⁻¹¹

Tabulka 9: Intervaly spolehlivosti model 5

Proměnná	Odhad	2,5 %	97,5 %
Konstanta	55,1424	45,5900	67,1476
Vzdálenost	-0,0159	-0,0188	-0,0130
Pohár	0,2395	0,1455	0,3433
Ročník	0,0081	0,0054	0,0108
Den	0,4251	0,2722	0,5950
Dopoledne	-0,3655	0,5203	0,7686
Odpoledne	-0,3158	-0,4339	-0,1786
Cena	0,0027	0,0024	0,0029
Krajské město	0,2955	0,1873	0,4146
Ocenění	0,5947	0,3845	0,8433
Google	0,0018	0,0015	0,0021

3.2 Interpretace výsledků

V této části je uvedena interpretace výsledků na základě modelu 5.

Vzdálenost je proměnná, která má negativní vliv na počet účastníků. V průměru se jedná o pokles necelých 2 %. Je to způsobeno již zmiňovaným faktem, že na uběhnutí závodů do 10 km nepotřebuje běžec trénovat tak, jako kdyby chtěl zaběhnout maraton. Kratších závodů se běžci účastní i několikrát do měsíce, zatímco maratonů a ultramaratonů jednou, je zde i delší regenerace.

Proměnná **pohár** má naopak pozitivní efekt, i toto se shoduje s očekáváním, neboť běžci v těchto závodech mohou závodit nejen o vítězství na jednotlivých závodech, ale také v celém poháru.

Proměnná **dětské závody** poukazovala na negativní vliv na závisle proměnnou. Tato proměnná byla již několikrát zmiňována, jelikož nebyla statisticky významná ani v jednom z uvedených modelů. Pořádání přidružených závodů neznamená zvýšení počtu běžců na závodech.

Další proměnnou, u které byl očekáván pozitivní vliv, byl **ročník**. Již v sekci Data bylo zmíněno, že s každým dalším rokem se dá očekávat více běžců na závodech. Jedním z faktorů může být tradice, že lidé tento běh běhají každým rokem. Dalším důvodem pak mohou být pozitivní ohlasy. Z těchto důvodů by se možná očekával vyšší nárůst účastníků. V průměru, ale s každým dalším ročníkem, můžeme očekávat nárůst o pouhých 4-5 závodníků.

U proměnné **den**, kde hodnota 1 značí, že závod je konán o víkendu nebo o státním svátku, byl také očekáván pozitivní vliv, jelikož většina závodů je konána v těchto dnech a lidé mají v tuto dobu i více volna. Odhad u této proměnné neposkytuje nejpřesnější informace, z důvodu šířky intervalu spolehlivosti, jedná se o průměrný nárůst mezi 27 % až 60 % běžců.

U kategoriální proměnné **čas**, byla jako referenční kategorie zvolena kategorie večer. Většina závodů je odstartována dopoledne nebo odpoledne, přesto pokud je závod odstartován dopoledne (respektive odpoledne) je účast o 37 % (resp. 32 %) nižší než u závodů startujících ve večerních hodinách. Večer jsou často startovány závody pro charitu, ať se jedná například o známý závod Běh pro Světlušku nebo závody se speciálními efekty jakým je Neon run pro Kapku naděje, který je zároveň i charitativním během. V tuto dobu startují i velmi populární Night run závody, u těchto závodů se dá očekávat velká účast. Lidé tyto závody většinou neběhají na čas, ale pro zábavu.

U proměnné **cena** by se očekával negativní vliv na účast. Ve všech modelech byl však odhadnut pozitivní vliv. Tento fakt může být způsoben reklamou, která ve výše odhadnutých modelech nebyla zcela zachycena. Známé závody mají vyšší startovné než závody, pořádané spolky mnohokrát i za dobrovolné startovné. U některých závodů dělá mnoho i samotné jméno závodu, ať se jedná o pražský maraton nebo o charitativní běhy. S velikostí startovného je někdy spojen i určitý komfort, který běžci očekávají, například přítomnost časomíry, občerstvovacích stanic nebo peněžní odměny pro vítěze či medaile pro každého, který proběhne cílem.

Krajské město, tato proměnná, má také pozitivní vliv na závisle proměnnou, toto není nikterak překvapivé. V krajských městech žije více lidí, tím i potencionálně více běžců, kteří by se mohli daného závodu účastnit. Velká část charitativních běhů je také pořádána v těchto městech.

Proměnná **ocenění** má stejně jako předchozí proměnné pozitivní vliv na počet účastníků. Závody oceněné za rok 2017, se mohli těšit velké účasti i v roce 2018. Ocenění většinou obdrží závody s největší účastí, proto jsou každým rokem oceňováni, až na výjimky, stejní organizátoři. Odhad u této proměnné je nejméně přesným, z důvodu nejširšího intervalu spolehlivosti.

Google byla další proměnná ze skupiny reklama. U této proměnné byl očekáván pozitivní vliv. Počet stránek vyhledaných na Googlu měl zachytit vliv reklamy neboli jak snadné je se o konkrétním závodě dozvědět. Odhad byl ovlivněn formou zadávání. Tato forma byla zvolena z již zmíněných důvodů, aby nedocházelo k záměně závodů nesoucích stejné jméno.

3.3 Testování hypotéz

V úvodu byly stanoveny výzkumné hypotézy. V empirické analýze došlo k jejich otestování a v této kapitole budou shrnuty výsledky těchto hypotéz.

H1: Závodníci si vybírají závody pouze na základě parametrů závodu.

Na základě parametrů závodu se nepodařilo vysvětlit průměrnou účast běžců na běžeckých závodech, jelikož model 1, nebyl zcela správně specifikován. Model neobsahoval proměnnou, které by se podařilo vysvětlit velké rozdíly v účasti na jednotlivých závodech. Další model byl proto doplněn o dvě proměnné ze skupiny reklama, jednalo se o proměnné ocenění a Google.

H2: Konání přidružených akcí přiláká na hlavní závod více běžců.

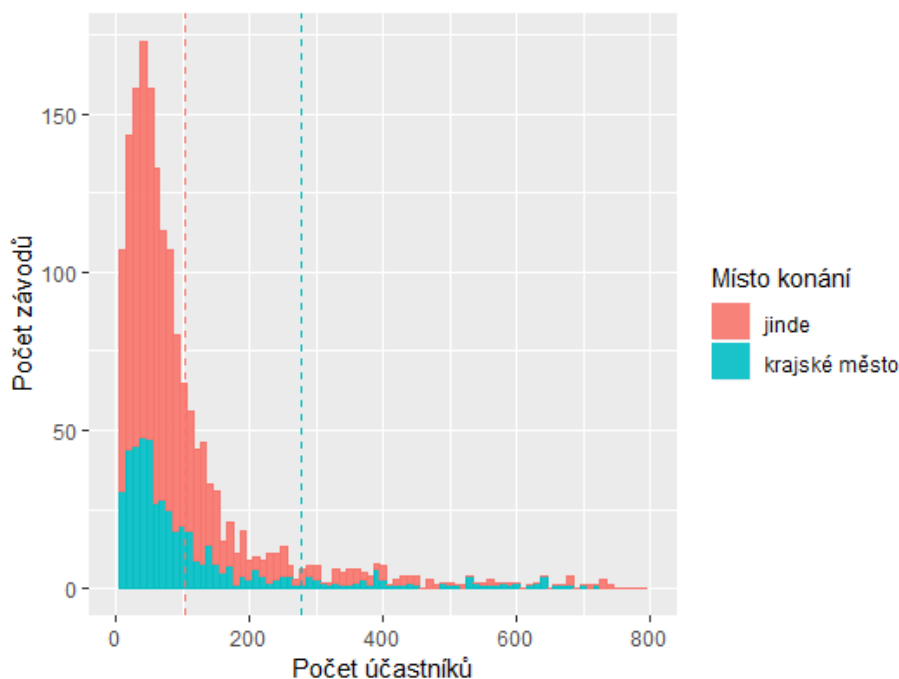
Proměnná dětský závod, která označuje doplňkové akce konané společně s hlavním závodem, se ukázala jako statisticky nevýznamná. Na tomto základě zamítáme hypotézu o tom, že by přidružené závody přilákaly více běžců na start. Toto zjištění se na první pohled může jevit jako překvapivé, ale pouze u 55 % závodů byl konán i přidružený závod. Průměrná účast u těchto závodů byla 136 běžců, zatímco u závodů bez přidružené akce 180 běžců (Tabulka2).

H3: Vysoké startovné negativně ovlivňuje počet účastníků.

Z důvodu kladného koeficientu u proměnné cena, byla i tato hypotéza zamítnuta. Jak již bylo zmíněno v kapitole Interpretace výsledků, vyšší cena může být spojována s větším komfortem, který běžci očekávají. Dalším faktorem může být, že závodů pořádaných například skupinou RunCzech, se účastní i světoví běžci. Na těchto závodech je jediná možnost v České republice, kde mohou i obyčejní běžci poměřit síly se světovou běžeckou elitou.

H4: Běžci dávají přednost závodům v krajských městech než v jiných lokalitách.

V grafu (Obrázek8) jsou vyobrazeny počty účastníků podle místa konání. Z důvodu lepší přehlednosti je zobrazen jen redukovaný počet běžců. Opět je vynecháno 43 závodů, ze kterých bylo 11 závodů konaných mimo krajské město. Přerušované čáry v grafu znázorňují průměry. U závodů konaných v krajském městě je průměrná návštěvnost 276 závodníků a u závodů konaných na jiných místech se účastní v průměru 104 běžců. Proměnná krajské město se jeví jako statisticky významná a má pozitivní vliv na závisle proměnnou. Z těchto důvodů není tato hypotéza zamítnuta.



Obrázek 8: Histogram počtu účastníků v závislosti na místě konání (prerušované čáry označují skupinové průměry)

3.4 Diskuze

Výsledky z této práce by mohly být využity organizátory běžeckých závodů. Na základě výsledků se mohou rozhodnout jakou vzdálenost zvolit, kde a kdy daný závod nebo seriál závodů uspořádat. Popřípadě jakou účast mohou očekávat v dalších ročnících.

Do budoucna by bylo vhodné lépe zachytit vliv reklamy. U proměnné Google měl velký vliv formát, který byl pro vyhledávání zvolen. Místo vyhledávání na Googlu by mohlo být zvoleno vyhledávání na Facebooku, a zde zaznamenávat počet sledujících. Mnoho organizátorů využívá Facebook jako hlavní komunikační kanál. Jedná se o bezplatnou formu reklamy. Zároveň se jedná o místo, kde si běžci přeprořádávají startovné, nebo organizátoři nabízejí startovné zadarmo formou soutěže. Další možnou proměnnou by mohla být částka vynaložená na propagaci závodu, i kdyby byla tato data dostupná mohl by nastat problém s velkými závody, které mají mnoho sponzorů, kteří jim poskytují reklamu zadarmo. Ne každý organizátor si může dovolit reklamu na dopravním prostředku v hlavním městě, zároveň také ne každý organizátor dokáže pozastavit dopravu ve městech, z důvodu konání závodu. Dále by mohlo být využito rozdílu mezi přihlášenými závodníky neboli startovní listinou a běžci, kteří opravdu na start přišli, tedy výslednou listinou. Tento rozdíl však nemusí být zajímavý pro organizátory, a to z důvodu nevratného startovného. Vhodnou binární proměnnou by mohl být též charitativní závod, který byl v této práci zmíněn v souvislosti s vyšší návštěvností, avšak tato proměnná nebyla zaznamenána a není tedy jasné, zda má i tato proměnná na počet účastníků vliv. V této práci byla proměnná den zvolena jako binární proměnná. Mohla by však být zvolena jako kategoriální, kde by kategorie tvořily jednotlivé dny v týdnu. Při sbírání dat bylo zjištěno, že nejméně závodů je pořádaných v pondělí jednalo se o pouhých 28 závodů. Toto může být spojeno s tím, že

většina závodů je konaných o víkendu a běžci si mohou nechávat pondělí jako regenerační den. Data jsou sesbírána pouze za jeden rok, do budoucna by bylo vhodné sesbírat data za více let a sledovat možnou sezónnost.

Závěr

Cílem této práce bylo zjistit, jaké faktory ovlivňují účast závodníků na běžeckých závodech. V první části byla představena data. Z důvodu neexistence vhodného datového souboru, bylo nejdříve nutné si samotná data obstarat. Jako výchozí bod posloužila běžecká termínovka, kde u každého závodu byla uvedena vzdálenost, místo a den konání a popřípadě zda je závod součástí poháru. U některých závodů byla uvedena také cena a počet účastníků. Zbylá data byla dohledávána individuálně pro každý závod z oficiálních webových stránek závodů, nebo portálů uvádějící výsledky závodů.

V následující části byla stručně popsána metodologie. Přesněji Poissonova regrese a problém s případným výskytem nadměrné disperze. Jako možnost řešení tohoto problému, zde byla uvedena negativně binomická regrese (NB2).

Poslední část se věnovala sestavením ekonometrického modelu. Prvně byl odhadnut model Poissonovou regresí, který obsahoval skupinu proměnných nazvanou parametry závodu. Tato regrese má předpoklad o rovnosti podmíněné střední hodnoty a rozptylu, který nebývá často splněn, ani tato data nebyla výjimkou. Z tohoto důvodu byl další model odhadnut negativně binomickou regresí. Stejně, jako tomu bylo v předchozím modelu, obsahoval i tento model pouze parametry závodu. Tento model byl zamítnut na základě reziduí, která poukazovala na možnost výskytu odlehlých pozorování. Do modelu 3 byla přidána druhá skupina proměnných nazvaná reklama. Od těchto proměnných se očekávalo vysvětlení rozdílů mezi více a méně navštěvovanými závody. Obě proměnné byly na 5 % hladině významnosti statisticky významné. Je tedy patrné, že i ve sportu má určitý vliv reklama a prestiž jednotlivých závodů. Proměnná dětský závod se jevila ve všech odhadnutých modelech jako statisticky nevýznamná, v posledním modelu již nebyla obsažena. U většiny proměnných vyšly vlivy na účast závodníků podle očekávání, až na proměnné cena a čas. U proměnné cena vyšel pozitivní vliv. Vyšší startovné je často spojováno s větším komfortem. U proměnné čas byla očekávána největší účast u kategorie dopoledne, protože v tuto dobu startuje nejvíce závodů. Večerní závody mají však svá specifika, není jich mnoho a mají zvláštní atmosféru. Závody patřící mezi ty navštěvovanější, jsou většinou pořádány v některém z krajských měst, nebo bývají součástí poháru.

Takto sestavený model by mohl být využit organizátory běžeckých závodů a napomoci k uspořádání závodu s nemalou účastí. Model by se dal vylepšit přidáním proměnné, která by lépe zachytila vliv reklamy. Po sesbírání dat za více let, by mohl být sledován například možný výskyt sezónnosti.

Použitá literatura

- [1] BEHEJ.COM s.r.o. Termínovka. *Běhej.com*. [online]. Praha: Běhej.com s.r.o., © 2017- [vid. 2019-3-31]. Dostupné z: <https://www.behej.com/terminovka>
- [2] CAMERON, A C., TRIVEDI, P K. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*. January 1986, vol. 1, no. 1, s. 29-53. ISSN 1099-1255.
- [3] CAMERON, A C., TRIVEDI, P K. 2013. *Regression analysis of count data*. 2. vyd. New York: Cambridge University Press. ISBN 978-1-107-66727-3.
- [4] CAMERON, A C., TRIVEDI, P K. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*. December 1990, vol. 46, no. 3, s. 347-364. ISSN 0304-4076.
- [5] ČESKÝ ATLETICKÝ SVAZ. Termínovka. *Český běh.cz*. [online]. Praha: eSports.cz s.r.o., © 2015- [vid. 2019-3-31]. Dostupné z: <https://www.ceskybeh.cz/term>
- [6] HILBE, J M. 2011. *Negative binomial regression*. 2. vyd. New York: Cambridge University Press. ISBN 978-0521-19815-8.
- [7] STOPNITO s.r.o. Závody. *STOPNITO*. [online]. Praha: STOPNITO s.r.o., © 2018- [vid. 2019-3-31]. Dostupné z: <https://www.stopnito.cz/zavo>

