Provozně-ekonomická fakulta

Hodnocení návštěvnosti ZOO Praha v závislosti na počasí

www.cz-milka.net
www.i-page.net

1. Obsah

2. Úvod

2.1 Téma projektu

Obliba ZOO Praha v posledních letech prudce stoupá. Nárůst počtu návštěvníků byl úzce spojen s povodněmi, které zoologickou zahradu těžce postihly roku 2002, nepochybně ale závisel také na aktuálním počasí.

Jako téma práce jsme si proto vybrali „Hodnocení návštěvnosti ZOO Praha v závislosti na počasí“, abychom si ověřili domněnku, že teplejší počasí přiměje k návštěvě zoologické zahrady větší počet návštěvníků.

2.2 Cíl projektu

Cílem naší práce je zjistit, zda návštěvnost ZOO Praha v jednotlivých měsících roku 2004 a 2005 závisí na průměrných měsíčních teplotách v dané lokalitě.

Problematiku jsme se rozhodli řešit pomocí korelační a regresní analýzy. Tato metoda se zabývá zkoumáním statistické závislosti, což nejlépe odpovídá našemu problému.

3. Hodnocené ukazatele

3.1 Vstupní data

Prvním krokem naší práce bylo získání vstupních dat, a to:

- počtu návštěvníků v jednotlivých měsících roku 2004 a 2005 a
- průměrné měsíční teploty v dané lokalitě v jednotlivých měsících roku 2004 a 2005.

Jelikož ZOO Praha na rozdíl od jiných zoologických zahrad neuvádí ve svých Výročních zprávách přesná čísla návštěvnosti, požádali jsme o tyto údaje vedení zoologické zahrady. Díky náměstku ÚKV Dr. Petru Schonfeldovi se nám podařilo získat údaje o návštěvnosti v jednotlivých měsících roku 2004 a 2005.

Teplotu v jednotlivých měsících roku 2004 a 2005 jsme získali z Internetu. Jedná se o průměrnou měsíční teplotu na území Prahy. Údaje o průměrné teplotě i počtu návštěvníků uvádíme v tabulce (Tab. 1).

Období	Počet návštěvníků [1]	Průměrná teplota ve °C [2]
Leden 2004	11 854	–3,4
Únor 2004	34 885	1,7
Březen 2004	39 089	3,5
Duben 2004	91 811	9,4
Květen 2004	106 507	11,8
Červen 2004	127 074	15,6
Červenec 2004	155 324	17,6
Srpen 2004	147 024	19,0
Září 2004	85 881	13,8
Říjen 2004	68 275	9,4
Listopad 2004	23 414	3,8
Prosinec 2004	80 745	–0,1
Leden 2005	58 630	0,8
Únor 2005	31 935	–3,1
Březen 2005	89 305	1,9
Duben 2005	121 310	9,9
Květen 2005	128 959	13,7
Červen 2005	137 990	16,7
Červenec 2005	156 877	18,5
Srpen 2005	207 449	16,8
Září 2005	113 195	15,0
Říjen 2005	115 869	9,7
Listopad 2005	28 077	2,6
Prosinec 2005	25 331	–0,4

Tab. 1 – Návštěvnost ZOO Praha a průměrná teplota v jednotlivých měsících roku 2004 a 2005.

3.2 Korelační a regresní analýza

Úlohu jsme se rozhodli řešit pomocí korelační a regresní analýzy. Tuto metodu jsme zvolili proto, že se zabývá zkoumáním statistické závislosti, která nejlépe odpovídá našemu problému.

Pomocí korelační a regresní analýzy je možné statistiky analyzovat vztahy mezi několika veličinami.

Tyto veličiny jsou často dvě:

- Nezávisle proměnná X, tzv. vysvětlující proměnná
- Závisle proměnná Y, tzv. vysvětlovaná proměnná

V našem případě je nezávisle proměnnou X průměrná měsíční teplota. Závisle proměnnou Y je pak počet návštěvníků ZOO, protože ten závisí na počasí.

4. Výpočet

4.1 SEMMA

Při řešení podobných příkladů je doporučeno držet se stanoveného postupu. Jedná se o pět etap statistické analýzy, která se zkráceně nazývá SEMMA.

Každé písmeno této zkratky má svůj význam:

- S – Sample
- E – Explore
- M – Modify
- M – Model
- A – Assess

4.2 Sample

V první etapě je nutné pořídit si výběrový soubor. Ten jsme uvedli ve 3. kapitole „Hodnocené ukazatele“.

Pokud se jedná o velkou databázi údajů, je doporučeno zvolit pouze náhodný výběr. V našem případě však nebylo nutné k tomuto kroku přistoupit.

4.3 Explore

Druhá etapa prozkoumává vlastnosti vybraného souboru a také jeho zvláštnosti pomocí různých charakteristik:

- Charakteristiky polohy – aritmetický průměr, medián, modus.
- Charakteristiky variability – rozptyl, směrodatná odchylka, variační rozpětí, variační koeficient.
- Netypické hodnoty – odlehlá pozorování.

4.3.1 Charakteristiky nezávislé proměnné

Pro základní charakteristiky proměnné teplota byly použity tabulky Moments (Tab. 2) a Quantiles (Tab. 3). Lze z nich určit, že ve sledovaném období byla minimální teplota –3,4°C a maximální teplota 19,0°C. Tabulky také znázorňují hodnotu 1. kvartilu, která je 1,8, a hodnotu 3. kvartilu, která dosahuje výše 15,3. Šikmost (skewness) hodnot teploty je –0,1312 a značí, že rozdělení četností je zešikmeno vpravo. Špičatost (kurtosis) má hodnotu –1,4558 a znamená, že rozdělení pozorovaných hodnot je plošší než rozdělení normální.

Moments

N	24.0000	Sum Wgts	24.0000
Mean	8.5083	Sum	204.2000
Std Dev	7.3816	Variance	54.4878
Skewness	-0.1312	Kurtosis	-1.4558
USS	2990.6200	CSS	1253.2183
CV	86.7571	Std Mean	1.5068

Quantiles

100% Max	19.0000	99.0%	19.0000
75% Q3	15.3000	97.5%	19.0000
50% Med	9.5500	95.0%	18.5000
25% Q1	1.8000	90.0%	17.6000
0% Min	-3.4000	10.0%	-0.4000
Range	22.4000	5.0%	-3.1000
Q3-Q1	13.5000	2.5%	-3.4000
Mode	9.4000	1.0%	-3.4000

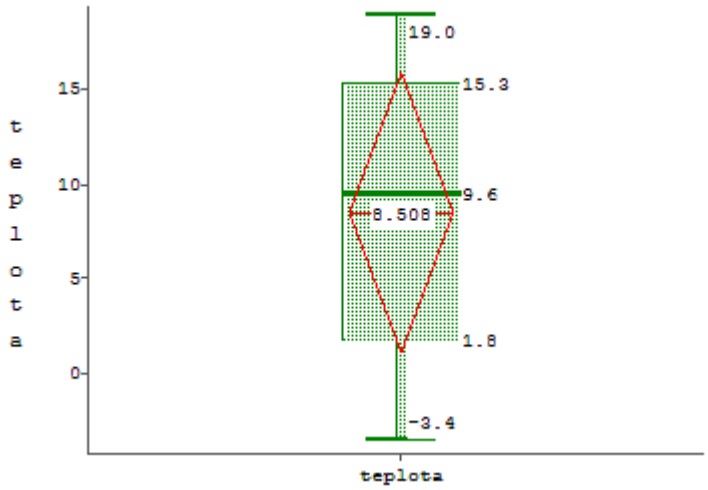
Tab. 2 – Tabulka Moments
pro nezávislou proměnnou teplota.

Tab. 3 – Tabulka Quantiles
pro nezávislou proměnnou teplota.

V tabulkách jsou uvedeny základní charakteristiky polohy. Jednou z nich je průměr (mean), jehož hodnota je 8,5083. Prostřední hodnota seřazeného souboru nazvaná medián, je 9,55, a nejčetnější hodnota daného souboru, kterou je takzvaný modus, je 9,4.

Obě tabulky uvádějí také charakteristiky variability. Rozptyl (variance) má hodnotu 54,4878 a vyjadřuje rozptýlení hodnot okolo charakteristické polohy. Odmocninou rozptylu je směrodatná odchylka (std dev), která dosahuje výše 7,3816. Dalšími charakteristikami variability je variační rozpětí (range), jehož hodnota je 22,4, a variační koeficient (coeficient of variance, CV), který dosahuje výše 86,7571.

Minimální a maximální hodnotu, průměr, medián a hodnoty 1. a 3. kvartilu zobrazuje také graf Box Plot (Obr. 1). Pomocí tohoto grafu lze znázornit také netypické hodnoty, které představují odlehlá pozorování. V našem případě však graf takové hodnoty nezobrazuje, z čehož vyplývá, že náš soubor nemá netypické hodnoty.



Obr. 1 – Box Plot nezávisle proměnné teplota.

4.3.2 Charakteristiky závislé proměnné

Postup pro sledování základních charakteristik proměnné návštěvníci je stejný, jako v případě proměnné teplota. K získání základních údajů jsme použili tabulku Moments (Tab. 4) a Quantiles (Tab. 5). Minimální počet návštěvníků ve sledovaném období byl 11.854 a maximální 207.449. Hodnota 1. kvartilu je 36.987 a hodnota 3. kvartilu 128.016,5. Lze určit také šikmost (skewness), jejíž hodnota je 0,2148, a špičatost (kurtosis) –0,6495.

Moments

N	24.0000	Sum Wgts	24.0000
Mean	91117.0833	Sum	2186810.00
Std Dev	52141.8437	Variance	2718771865
Skewness	0.2148	Kurtosis	-0.6495
USS	2.618E+11	CSS	6.2532E+10
CV	57.2251	Std Mean	10643.4093

Tab. 4 – Tabulka Moments
pro závislou proměnnou návštěvníci.

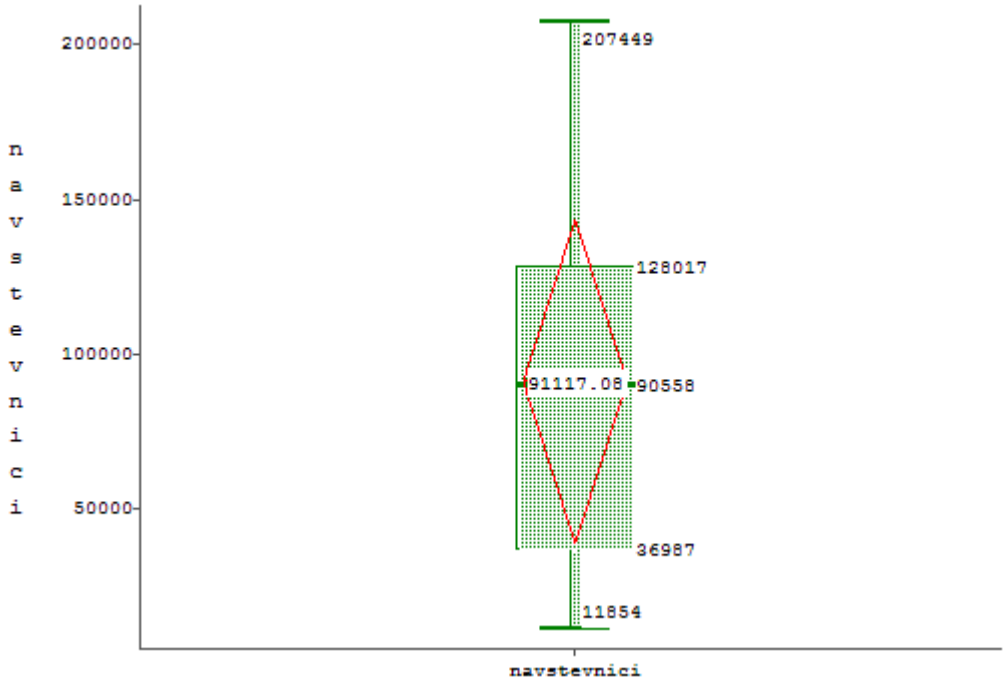
Quantiles

100% Max	207449.000	99.0%	207449.000
75% Q3	128016.500	97.5%	207449.000
50% Med	90558.0000	95.0%	156877.000
25% Q1	36987.0000	90.0%	155324.000
0% Min	11854.0000	10.0%	25331.0000
Range	195595.000	5.0%	23414.0000
Q3-Q1	91029.5000	2.5%	11854.0000
Mode	.	1.0%	11854.0000

Tab. 5 – Tabulka Quantiles
pro závislou proměnnou návštěvníci.

Průměrný počet návštěvníků ve sledovaném období byl 91.117,0833. Hodnota mediánu je 90.558. Jelikož žádná hodnota počtu návštěvníků se neopakuje, není možné určit hodnotu modus. Rozptyl hodnot je 2.718.771.865. Odmocninou rozptylu je pak směrodatná odchylka, jejíž hodnota je 52.141,8437. Variační rozpětí je 195.595 a variační koeficient má hodnotu 57,2251.

Pro úplnost předkládáme také graf Box Plot (Obr. 2), který ani v tomto případě neukazuje odlehlá pozorování v podobě netypických hodnot.



Obr. 2 – Box Plot závisle proměnné návštěvníci.

4.4 Modify

Ve třetím kroku postupu SEMMA je nutno zkontrolovat správnost dat a případně provést jejich modifikaci. Ta spočívá v úpravě souboru, který nemá vhodné vlastnosti. Případnou nevhodnost odhalíme ve druhé fázi Explore.

Jelikož Box Plot neukázal v případě závisle ani nezávisle proměnné na netypické hodnoty, lze z toho usuzovat, že naše vstupní data jsou bez odchylek a extrémů.

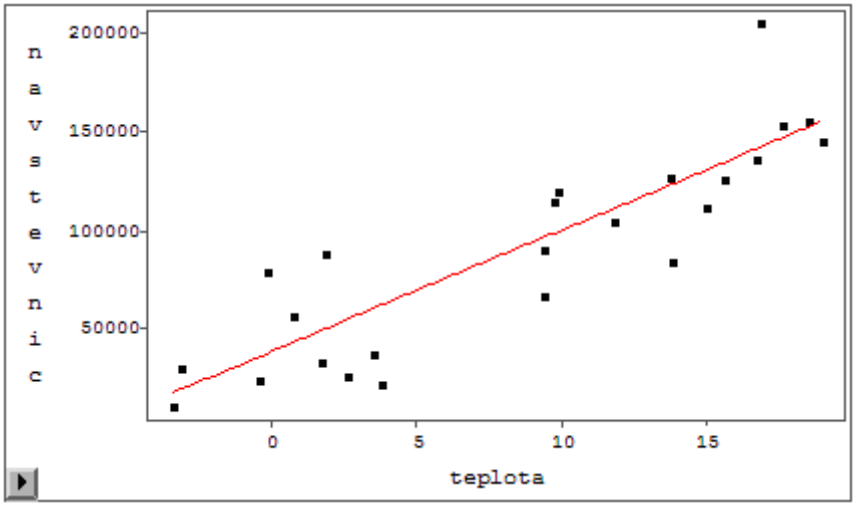
4.5 Model

Čtvrtý krok SEMMY slouží ke zvolení příslušné modelovací procedury, díky které dostaneme konečné výsledky.

Nejprve určíme regresní přímku, kterou lze vyčíslit (Tab. 6) a graficky zobrazit (Obr. 3) pomocí programu SAS. Přímka má tvar Y = 38.509,9 + 6.183,02X. Graf poukazuje na jedno odlehlé pozorování.

Model Equation				
navstevnici	=	38509.9	+	6183.02 teplota

Tab. 6 – Regresní přímka.



Obr. 3 – Regresní přímka.

Hodnota α, tedy chyba 1. druhu nazývaná také hladina významnosti, je 0,05. Z modelu regrese (Tab. 7) lze zjistit, přepočítaná hladina významnosti je menší než 0,0001 a tedy α je větší než přepočítaná hodnota. Z toho vyplývá, že model je statisticky významný a lze o něm tedy říci, že je zobecnitelný.

Zároveň lze říci, že existuje statisticky významný rozdíl. Z tohoto důvodu zamítáme nulovou hypotézu H_0 a přijímáme alternativní hypotézu H_A.

Na základě koeficientu determinance (R-Square), jehož hodnota je 0,7662, lze říci, že ze 76,62% lze závislou proměnnou návštěvníci vysvětlit pomocí nezávislé proměnné teplota. Ze 76,62% tedy počet návštěvníků zoologické zahrady závisí na teplotě.

Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
———	1	1	4.791E+10	22	664612733	0.7662	72.09	<.0001

Tab. 7 – Model regrese.

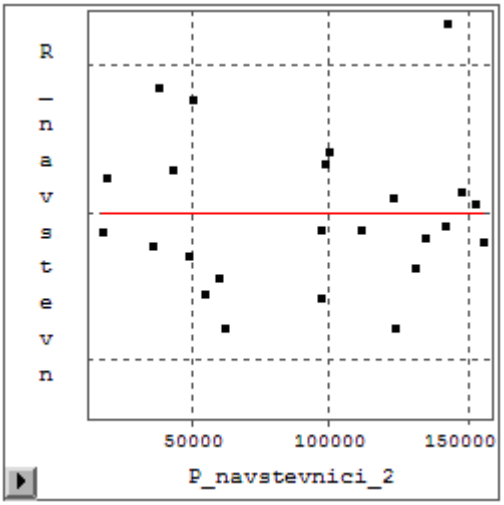
Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
Intercept	1	38509.8517	8129.1653	4.74	<.0001	.	0
teplota	1	6183.0243	728.2340	8.49	<.0001	1.0000	1.0000

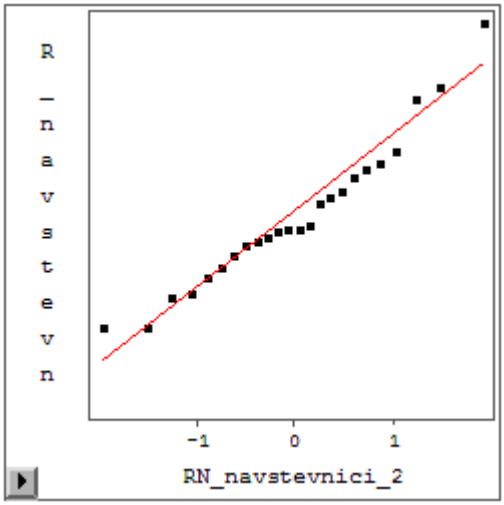
Tab. 8 – Odhad parametru.

V tabulce odhadu parametru (Tab. 8) se nachází hodnoty regresní konstanty a (38.509,8517) a koeficientu u-té regresní funkce b (6.183,0243), které se dosazují do základního tvaru regresní přímky Y = a + bX. Jak již bylo uvedeno výše, přímka má tvar Y = 38.509,9 + 6.183,02X.

Rezidua mají být nezávislá, náhodná, normální, mají mít konstantní rozptyl a střední hodnotu nula. Graf reziduí (Obr. 4) poukazuje na jednu odlehlou hodnotu. Při testu normality (Obr. 5) lze odhalit, že rozptyl je konstantní, s výjimkou dvou krajních hodnot.



Obr. 4 – Graf reziduí.



Obr. 5 – Test normality.

V dalším kroku přikročíme k programování, abychom odhalili problematické proměnné. Použijeme následující proceduru:

```
proc reg data=sasuser.projekt;
model navstevnici=teplota /r influence;
run;
```

Hvězdičky poukazují na problematická pozorování (Tab. 6). V našem případě lze říci, že takové pozorování má číslo 20.

Pro zjištění významných hodnot použijeme sloupec Hat Diag H (Tab. 7), který udává hodnotu z diagonály takzvané „kloboukové matice“. Pokud je tato hodnota Leverage větší, než $2 \times \frac{p}{n}$, kde p je počet regresních parametrů a n počet pozorování, lze říci, že hodnoty jsou významné. Po dosazení do vzorce zjistíme, že $2 \times \frac{2}{24} = \frac{4}{24} \doteq 0{,}17$. V našem případě však žádná hodnota není větší, než Leverage.

Dále je nutné zjistit, zda jsou pozorování vlivná. K tomu použijeme hodnoty ze sloupce Cook's D (Tab. 6), které udávají poměr vzdáleností X a Y. Spočítáme si limitní hodnotu, která je $\frac{4}{24} \doteq 0{,}17$ a zjistíme, že 12. pozorování jen těsně přesahuje tuto hodnotu (0,172) a lze ho tedy ještě považovat za nevlivné, ale 20. pozorování s hodnotou 0,377 limit výrazně přesahuje. Lze tedy říci, že toto pozorování je vlivné a negativně ovlivňuje regresní přímku. V takovém případě by bylo lepší hodnotu ze souboru vyřadit. V našem případě to však není možné, jelikož se jedná o roční hodnoty, které byly skutečně naměřeny.

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	11854	17488	10144	-5634	23701	-0.238	| |	0.005
2	34885	49021	7230	-14136	24745	-0.571	*| |	0.014
3	39089	60150	6403	-21061	24972	-0.843	*| |	0.023
4	91811	96630	5302	-4819	25229	-0.191	| |	0.001
5	106507	111470	5783	-4963	25123	-0.198	| |	0.001
6	127074	134965	7373	-7891	24703	-0.319	| |	0.005
7	155324	147331	8457	7993	24353	0.328	| |	0.006
8	147024	155987	9277	-8963	24053	-0.373	| |	0.010
9	85881	123836	6522	-37955	24941	-1.522	***| |	0.079
10	68275	96630	5302	-28355	25229	-1.124	**| |	0.028
11	23414	62005	6281	-38591	25003	-1.543	***| |	0.075
12	80745	37892	8185	42853	24446	1.753	|***	0.172
13	58630	43456	7694	15174	24605	0.617	|*	0.019
14	31935	19342	9958	12593	23779	0.530	|*	0.025
15	89305	50258	7131	39047	24774	1.576	|***	0.103
16	121310	99722	5359	21588	25217	0.856	|*	0.017
17	128959	123217	6480	5742	24952	0.230	| |	0.002
18	137990	141766	7955	-3776	24522	-0.154	| |	0.001
19	156877	152896	8980	3981	24166	0.165	| |	0.002
20	207449	142385	8010	65064	24504	2.655	|*****	0.377
21	113195	131255	7074	-18060	24791	-0.729	*| |	0.022
22	115869	98485	5333	17384	25222	0.689	|* |	0.011
23	28077	54586	6797	-26509	24868	-1.066	**| |	0.042
24	25331	36037	8353	-10706	24389	-0.439	| |	0.011

Obr. 6 – Výstupní hodnoty.

Obs	RStudent	Hat Diag H	Cov Ratio	DFFITS	-------DFBETAS------- Intercept	teplota
1	-0.2325	0.1548	1.2919	-0.0995	-0.0983	0.0851
2	-0.5623	0.0787	1.1561	-0.1643	-0.1633	0.1127
3	-0.8377	0.0617	1.0952	-0.2148	-0.2075	0.1223
4	-0.1868	0.0423	1.1422	-0.0393	-0.0216	-0.0048
5	-0.1932	0.0503	1.1516	-0.0445	-0.0121	-0.0184
6	-0.3128	0.0818	1.1842	-0.0934	0.0067	-0.0654
7	0.3214	0.1076	1.2179	0.1116	-0.0216	0.0874
8	-0.3652	0.1295	1.2449	-0.1409	0.0367	-0.1160
9	-1.5718	0.0640	0.9387	-0.4110	-0.0296	-0.2429
10	-1.1310	0.0423	1.0182	-0.2377	-0.1305	-0.0291
11	-1.5969	0.0594	0.9278	-0.4011	-0.3845	0.2190
12	1.8465	0.1008	0.9034	0.6182	0.6182	-0.4735
13	0.6078	0.0891	1.1635	0.1901	0.1898	-0.1387
14	0.5207	0.1492	1.2573	0.2180	0.2157	-0.1851
15	1.6350	0.0765	0.9352	0.4706	0.4669	-0.3176
16	0.8507	0.0432	1.0719	0.1808	0.0889	0.0342
17	0.2251	0.0632	1.1659	0.0585	0.0047	0.0341
18	-0.1505	0.0952	1.2104	-0.0488	0.0070	-0.0366
19	0.1611	0.1213	1.2460	0.0598	-0.0143	0.0485
20	3.1470	0.0965	0.5609	1.0286	-0.1536	0.7755
21	-0.7205	0.0753	1.1303	-0.2056	0.0057	-0.1374
22	0.6808	0.0428	1.0976	0.1440	0.0741	0.0234
23	-1.0695	0.0695	1.0608	-0.2923	-0.2875	0.1850
24	-0.4307	0.1050	1.2049	-0.1475	-0.1475	0.1146

Obr. 7 – Výstupní hodnoty.

4.6 Assess

V poslední etapě je nutno vyhodnotit výsledky a posoudit jejich správnost a kvalitu. V případě rozporu je nutno vrátit se zpět ke kroku Explore a postup opravit a opakovat.

Jelikož 12. ani 20. pozorování nelze z výše uvedených důvodů vyřadit ze souboru, je nutné zamítnout možnost hrubé chyby. V našem případě tedy působily další vlivy. Jak jsme dodatečně zjistili, 12. pozorování bylo ovlivněno akcí, která se konala v zoologické zahradě. Dne 29. listopadu 2004 byl slavnostně otevřen unikátní pavilon Indonéská džungle, který v prosinci téhož roku i přes nízké teploty nepochybně přilákal do zoologické zahrady větší počet návštěvníků.

Dvacáté pozorování bylo také ovlivněno akcemi, které zoologická zahrada pořádala. Po povodních roku 2002 probíhaly rozsáhlé renovace postižené části ZOO, které mimo jiné završily dne 7. srpna 2005 otevřením expozice lemurů a ostrova kotulů a dne 14. srpna téhož roku otevřením expozice Vodní svět a opičí ostrovy. Obě slavnostní akce, kterých se zúčastnily také známé osobnosti, nepochybně přilákaly do zoologické zahrady velký počet návštěvníků i přes neobvykle chladné počasí.

5. Závěr

Jak vyplynulo z hodnocení charakteristik polohy a variability ve fázi Explore, ani jedna proměnná nevykazuje netypické hodnoty v podobě odchylek a extrémních hodnot. O hodnoceném souboru lze říci, že má vhodné vlastnosti a je možné s ním dále pracovat.

Z regresní a korelační analýzy vyplývá, že model je statisticky významný a je tedy zobecnitelný. Mezi hodnotami existuje statisticky významný rozdíl, který vede k zamítnutí nulové hypotézy a přijetí hypotézy alternativní.

Téměř ze 77% závisí počet návštěvníků zoologické zahrady na teplotě. Lze také říci, že neexistují významné hodnoty, ale že 12. a 20. pozorování je vlivné a negativně ovlivňuje regresní přímku. Jedná se o naměřené hodnoty a proto není možné tato pozorování vyřadit ze souboru. Lze však zamítnout možnost hrubé chyby. Z dalšího zkoumání vyplynulo, že obě hodnoty byly ovlivněny akcemi konanými v areálu zoologické zahrady.

Návštěvnost zoologické zahrady v Praze je tedy závislá na teplotě. Lze však doporučit, aby ZOO v návštěvnicky slabších měsících pořádala více akcí pro návštěvníky, protože zvyšují návštěvnost.

6. Seznam literatury

[1] SCHONFELD, Petr. *FW: Návštěvnost v roce 2005.* [online] 18. října 2006 8:43. [cit. 2006-11-18]. Osobní komunikace.

[2] *Počasí.* [online] c2006. [cit. 2006-11-18]. Dostupné z <http://pocasi.divoch.cz>.

[3] KÁBA, Bohumil, SVATOŠOVÁ, Libuše. *Statistika.* ČZU 2005. 3. vydání, 3. dotisk. ISBN 80-213-0746-3.

[4] PRÁŠILOVÁ, Marie, SVATOŠOVÁ, Libuše. *Cvičení ze statistiky.* ČZU 2004. 4. vydání, 3. dotisk. ISBN 80-213-0712-9

7. Seznam obrázků a tabulek

7.1 Seznam obrázků

7.2 Seznam tabulek