Logistická regrese

Menu:	QCExpert	Regrese	Logistická

Modul Logistická regrese umožňuje analýzu dat, kdy odezva je binární, nebo frekvenční veličina vyjádřená hodnotami 0 nebo 1, případně poměry v intervalu < 0, 1 >. Poměry představují podíl pozitivních výsledků v případě více opakovaných měření při téže hodnotě nezávisle proměnné *x*. Počet měření by měl přitom být pro každý poměr přibližně stejný. Logistická regrese se používá při modelování pravděpodobnosti nějakého v závislosti na hodnotě spojité proměnné. Předpokládá se, že náhodná proměnná má binomické rozdělení s parametrem π, který odpovídá pravděpodobnosti výsledku „1“ a mění se monotónně s hodnotou nezávisle proměnné. Výsledný model je právě odhadem tohoto parametru v závislosti na *x*. Použití logistického modelu je velmi široké a zahrnuje řadu velmi rozdílných oborů. Typickými aplikacemi jsou v technologii odhad rizika selhání či poruchy za určitých podmínek, ve finančnictví predikce bonity klienta (např. rizika nesplácení úvěru) v závislosti na ekonomických ukazatelích, v biologii a ekologii pravděpodobnost úhynu organismu v závislosti na koncentraci toxických látek, v marketingu pravděpodobnost přechodu klienta ke konkurenci, v lékařství a farmacii modelování účinnosti léků a podobně.

Pravděpodobnost v závislosti na proměnné *x* je zde modelován pomocí logistického modelu

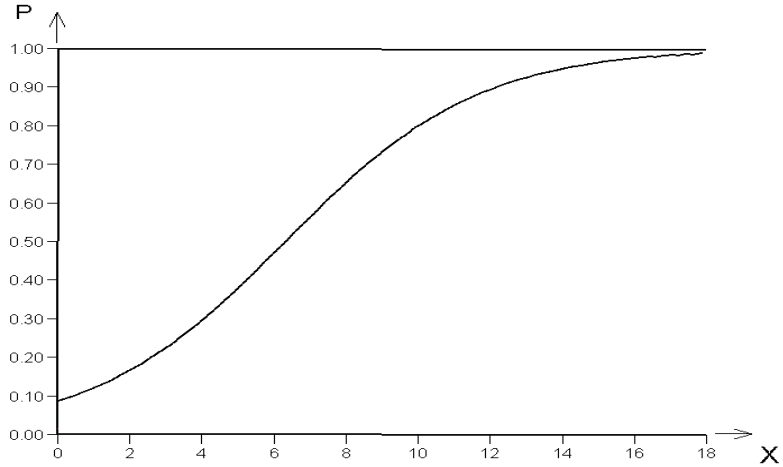
$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

neboli po úpravě

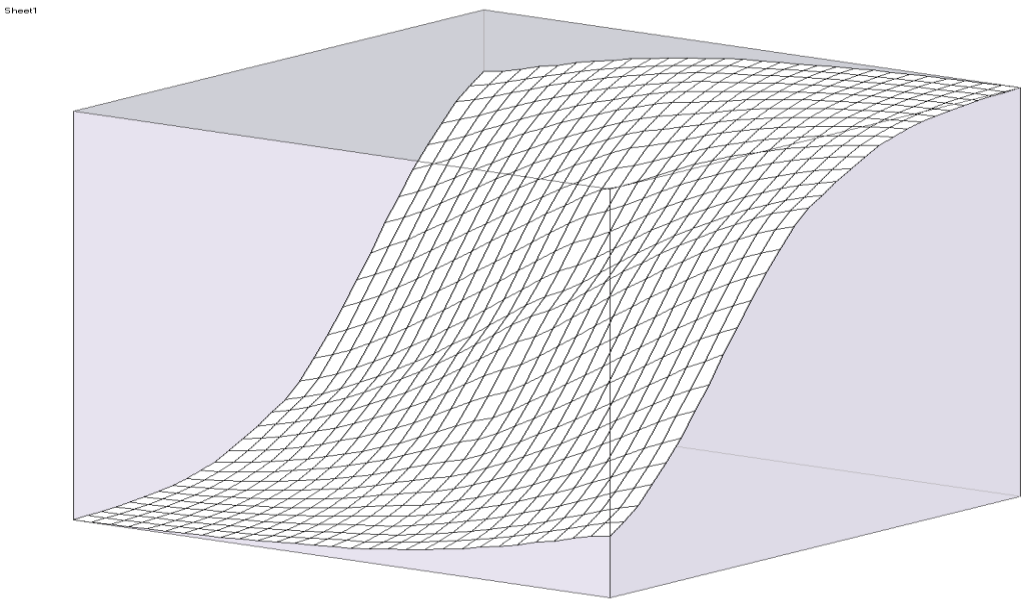
$$\log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x .$$

Výraz log ($\pi(x) / (1 - \pi(x))$) se nazývá logit. Hodnoty α a β jsou regresní koeficienty a k jejich odhadu *a*, *b* je použita iterativní metoda nejmenších čtverců. Touto metodou se získají maximálně věrohodné odhady α a β. Logistický regresní model lze vyjádřit sigmoidální křivkou $\pi(x)$ vyjadřující odhad závislosti pravděpodobnosti výskytu sledovaného jevu v závislosti na *x*. Tento model lze pak využít pro predikci pravděpodobnosti nebo rizika při nastavených hodnotách *x*. Nezávisle proměnná *x* může přitom být i vícerozměrná, $\mathbf{x} = (x_1, ..., x_m)$. Odpovídající model má pak tvar analogický lineární regresi.

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + + \beta_m x_m)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + + \beta_m x_m)} .$$



Obrázek 1 Logistický regresní model pro 1 proměnnou x



Obrázek 2 Ilustrační ukázka 3D logistického modelu se 2 proměnnými

Pomocí stanoveného modelu lze pak predikovat (předpovědět) pravděpodobnost nastání jevu při nových hodnotách nezávisle proměnných.

Data a parametry

Data musí zahrnovat jeden nebo více sloupců nezávisle proměnné a jeden sloupec závisle proměnné. Nezávisle proměnné, neboli prediktory mohou nabývat libovolných číselných hodnot. Binární závisle proměnná musí mít hodnoty 0 nebo 1, které odpovídají výskytu či nevýskytu sledovaného jevu (zlomení, porucha, úhyn, ztráta klienta, apod.). Přitom nezáleží na volbě, zda budeme výskyt označovat jedničkou a nevýskyt nulou, nebo naopak. Binární proměnná odpovídá případu, kdy pro danou hodnotu prediktoru máme pouze jediný výsledek typu ano-ne, viz Obrázek 3. Poměrná, neboli frekvenční závisle proměnná odpovídá případu, kdy pro danou hodnotu prediktoru x provedeme p testů a máme tedy p výsledků z nichž r je pozitivních a $p - r$ negativních. Do sloupce závisle proměnné pak můžeme zapsat poměr r/p, případně $(p - r)/p$, viz Obrázek 4. Nezávisle proměnná může být vícerozměrná – tedy ve více sloupcích, Obrázek 5.

Nezávisle proměnná	Binární závisle proměnná
Zátěž dílu	Porucha dílu
0	0
0	0
0	0
1	0
1	1
2	0
3	0
4	1
4	1
4	0
.....	.....

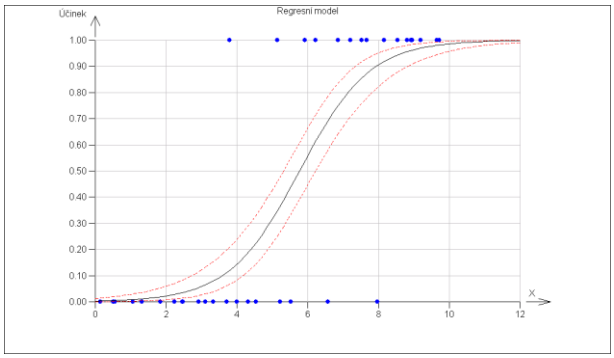
Obrázek 3 Data s binární závisle proměnnou

Nezávisle proměnná	Poměrná závisle proměnná
Koncentrace mg/l	Účinek (podíl z 10)
0	0
1	0
2	0.2
3	0.3
4	0.7
5	0.9
6	1

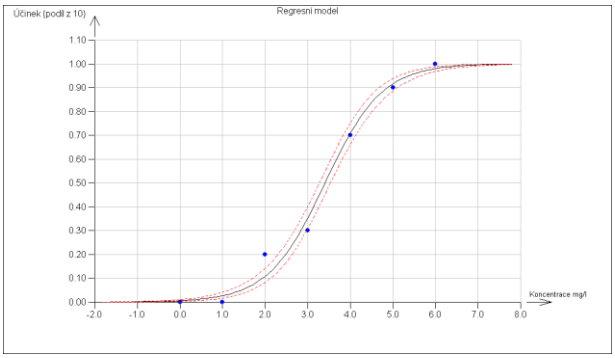
Obrázek 4 Data s poměrovou závisle proměnnou (poměr výskytů z 10 pokusných jednotek)

Nezávisle proměnné			Binární závisle proměnná
Doba	Intenzita	Frekvence	Porucha
8	34	200	1
5	38	250	0
5	35	250	0
7	40	200	1
6	29	100	1
4	35	150	0
11	37	150	1
4	28	200	0
8	32	200	0
6	30	250	1

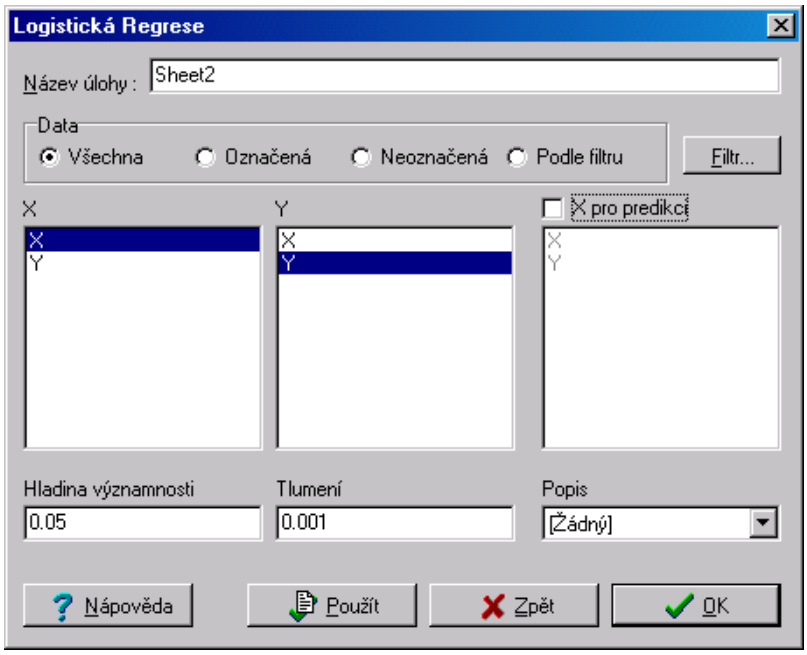
Obrázek 5 Data s více nezávisle proměnnými a binární závisle proměnnou



Obrázek 6 Logistický model pro binární data



Obrázek 7 Logistický model pro poměrná data



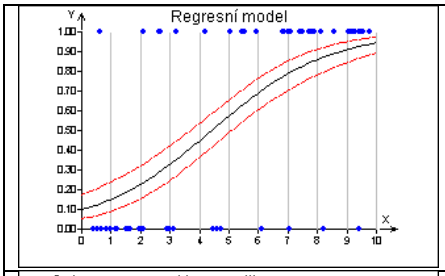
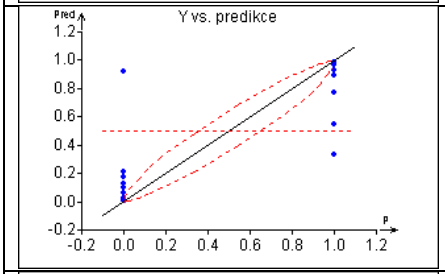
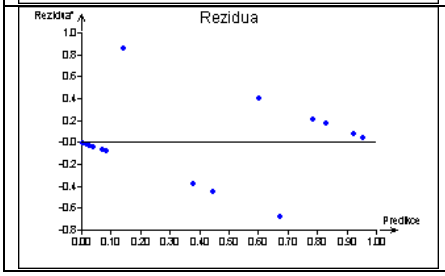
Obrázek 8 Dialogový panel modulu logistická regrese

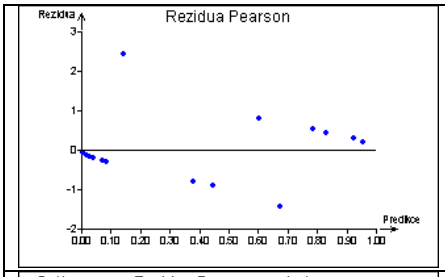
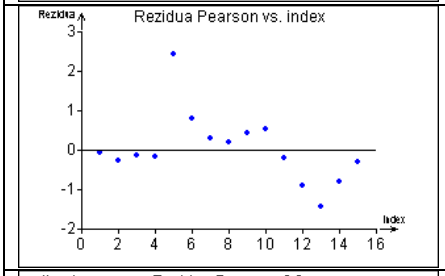
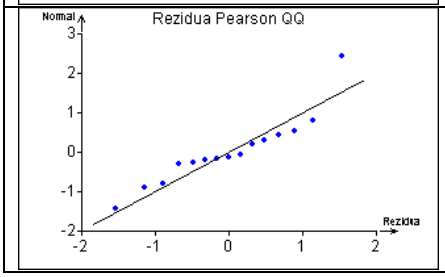
V dialogovém panelu logistické regrese zvolíme jednu nebo více závisle proměnných *X* a jednu nezávisle proměnnou *Y*. Máme-li hodnoty nezávisle proměnné, pro něž chceme zjistit odpovídající odezvu, tedy pravděpodobnost nastání jevu, označíme příslušné sloupce v poli *X pro predikci*. Počet nezávisle proměnných a proměnných pro predikci musí být stejný.

Protokol

Počet dat Nezávisle proměnné Závisle proměnná	Počet platných datových řádků Seznam nezávisle proměnných Název závisle proměnné
Počet iterací Max věrohodnost	Počet iterací výpočetního algoritmu při výpočtu parametrů metodou maximální věrohodnosti Přirozený logaritmus maximální dosažené věrohodnosti.
Odhady parametrů Parametr Odhad Sm. odch p-hodnota	Vypočítané hodnoty parametrů logistického modelu Název parametru u nezávisle proměnné, *Abs* je absolutní člen Odhad parametru Směrodatná odchylka parametru udává hladinu významnosti, při níž je statistická významnost parametru právě zamítnuta
Tabulka predikce Index *Název závisle proměnné* Predikce Spodní mez Horní mez	Tabulka hodnot pravděpodobnosti predikovaných logistickým modelem pro zadané hodnoty nezávisle proměnné Pořadové číslo Seznam nezávisle proměnných Pravděpodobnost pozitivní odezvy predikovaná modelem Spodní mez predikce při zadané hladině významnosti Horní mez predikce při zadané hladině významnosti

Grafy

	Průběh regresní závislosti s vyznačeným pásem spolehlivosti modelu a s naměřenými daty. Tento graf se zobrazí pouze při jediné nezávisle proměnné *x*. Pás spolehlivosti (spodní a horní červená křivka) vyznačuje pro každou hodnotu nezávisle proměnné interval. v němž lze očekávat na zadané hladině spolehlivosti 1 – α (obvykle 0.95, tedy 95%), že v něm leží skutečná pravděpodobnost nastání jevu *y* = 1.
	Graf pozorované odezvy *y* versus predikovaná pravděpodobnost. V případě dobrého proložení se soustřeďují body v grafu poblíž hodnot (0, 0) a (1, 1).
	Graf prostých reziduí, tj. vzdáleností dat od logistické křivky. Tato rezidua mají jinou povahu, než rezidua v lineární nebo nelineární regresi. Nemohou mít normální rozdělení a jejich hodnoty jsou vždy mezi -1 a 1.

	Transformovaná Pearsonova rezidua jsou srovnatelná s klasickými rezidui např. v lineární regresi. Mají normované normální rozdělení $N(\mu = 0, \sigma^2 = 1)$ a lze jich použít pro diagnostiku vybočujících měření.
	Transformovaná Pearsonova rezidua jsou srovnatelná s klasickými rezidui např. v lineární regresi. Mají normované normální rozdělení $N(\mu = 0, \sigma^2 = 1)$ a lze jich použít pro diagnostiku vybočujících měření. Pokud je pořadí dat neuspořádané podle predikrotu, může tento graf lépe prezentovat rozdělení reziduí, než graf předchozí.
	QQ-graf transformovaných reziduí je účinným diagnostickým nástrojem pro posouzení normality reziduí. Body, které se výrazně odchylují od přímky jsou podezřelé odlehlé body a je vhodné ověřit jejich validitu. V ideálním případě leží body přibližně na vyznačené přímce.