Hodnocení rizika pomocí regresních modelů a s využitím software

1 Regresní parabola

V tabulce jsou uvedeny počty přežívajících osob s diagnostikovaným nádorovým onemocněním kůže v populaci mužů kraje Praha (vysvětlovaná proměnná Y) od roku 1989 do roku 2003 (vysvětlující proměnná je čas a bude značena t a udaná v letech). Cílem je popsat dynamiku tohoto onemocnění.

Pozorování	i	1	2	3	4	5	6	7	8
Rok	t_i	1989	1990	1991	1992	1993	1994	1995	1996
Počet pacientů	Y_i	3477	3663	3825	3944	4189	4367	4633	4782
Pozorování	i	9	10	11	12	13	14	15	
Rok	t_i	1997	1998	1999	2000	2001	2002	2003	
Počet pacientů	Y_i	5090	5495	5893	6406	6801	7269	7832	

Abychom se vyhnuli numerické nestabilitě při výpočtech, budeme pracovat s vysvětlovanou proměnnou $X = t - 1988$, pak $x_i = t_i - 1988 = i,\ i = 1, 2, \ldots, 15$. Postupně dostaneme

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 15 & 120 & 1240 \\ 120 & 1240 & 14400 \\ 1240 & 14400 & 178312 \end{pmatrix},$$

$$\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0{,}7934 & -0{,}2044 & 0{,}0110 \\ -0{,}2044 & 0{,}0656 & -0{,}0039 \\ 0{,}0110 & -0{,}0039 & 0{,}0002 \end{pmatrix},$$

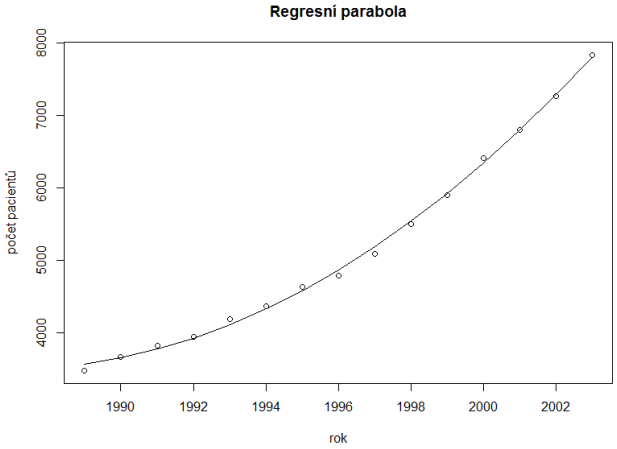
$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 77666 \\ 706002 \\ 7844260 \end{pmatrix} \text{ a } \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 3517{,}6703 \\ 34{,}4563 \\ 16{,}7469 \end{pmatrix}.$$

Regresní parabola má rovnici

$$\widehat{y} = 3517{,}67 + 34{,}46\,(t - 1988) + 16{,}75\,(t - 1988)^2.$$

```
data4<-read.table("nador_kuze.txt",header=T)
attach(data4)
names(data4)
plot(data4)
t<-1:15
m4<-lm(pacienti~t+I(t^2))
summary(m4)
confint(m4)
plot(rok,pacienti,ylab="počet pacientů", main="Regresní parabola")
lines(rok,predict(m4))
```

Residual standard error: 62.07 on 12 degrees of freedom Multiple R-squared: 0.9983, Adjusted R-squared: 0.998 F-statistic: 3474 on 2 and 12 DF, p-value: $<$ 2.2e-16



Obr. 1: Počty přežívajících osob s diagnostikovaným nádorovým onemocněním kůže v populaci mužů kraje Praha

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3517.6703	55.2859	63.63	0.0000
t	34.4563	15.9004	2.17	0.0511
I(t^2)	16.7469	0.9664	17.33	0.0000

2 Dva lineární regresory

V tabulce jsou uvedena data o počtu úmrtí v Londýně (hodnoty proměnné Y) od 1. do 15. 12. 1952, kdy Londýn postihla mimořádně silná mlha. Dále jsou uvedeny hodnoty proměnné X, která představuje průměrné znečištění vzduchu v County Hall uváděné v mg/m^3 a hodnoty proměnné Z, která představuje průměrný obsah oxidu siřičitého (počet částic na jeden milion). Cílem je popsat závislost počtu úmrtí Y na regresorech X a Z pomocí lineárního regresního modelu.

Den	1	2	3	4	5	6	7	8
Počet úmrtí	112	140	143	120	196	294	513	518
Znečištění vzduchu	0,3	0,49	0,61	0,49	2,64	3,45	4,46	4,46
Oxid siřičitý	0,09	0,16	0,22	0,14	0,75	0,86	1,34	1,34
Den	9	10	11	12	13	14	15	
Počet úmrtí	430	274	255	236	256	222	213	
Znečištění vzduchu	1,22	1,22	0,32	0,29	0,5	0,32	0,32	
Oxid siřičitý	0,47	0,47	0,22	0,23	0,26	0,16	0,16	

$$\mathbf{X'X} = \begin{pmatrix} 15{,}000 & 21{,}090 & 6{,}870 \\ 21{,}090 & 63{,}216 & 18{,}724 \\ 6{,}870 & 18{,}724 & 5{,}657 \end{pmatrix}, \quad \mathbf{X'Y} = \begin{pmatrix} 3922 \\ 7654{,}350 \\ 2439{,}540 \end{pmatrix},$$

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 89{,}5 \\ -220{,}3 \\ 1051{,}8 \end{pmatrix}$$

	2.5 %	97.5 %
(Intercept)	3397.2127	3638.1280
t	-0.1877	69.1003
I(t^2)	14.6414	18.8524

Tedy regresní rovnice vysvětlující počet úmrtí Y v závislosti na průměrném znečištění vzduchu x a průměrném obsahu z oxidu siřičitého je

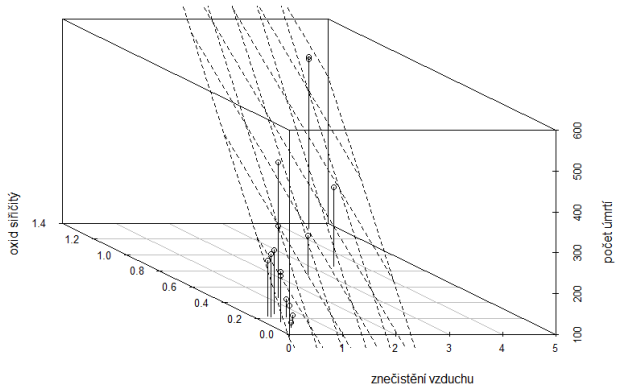
$$\widehat{y} = 89{,}511 - 220{,}324x + 1051{,}816z.$$

```
data5<-read.table("umrti_Londyn.txt",header=T)
attach(data5)
names(data5)
plot(data5)
m5<-lm(umrti~znecisteni+SO2)
summary(m5)
confint(m5)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	89.5108	25.0782	3.57	0.0039
znecisteni	-220.3244	58.1431	-3.79	0.0026
SO2	1051.8165	212.5960	4.95	0.0003

Residual standard error: 52.96 on 12 degrees of freedom Multiple R-squared: 0.859, Adjusted R-squared: 0.8355 F-statistic: 36.57 on 2 and 12 DF, p-value: 7.844e-06

	2.5 %	97.5 %
(Intercept)	34.8700	144.1516
znecisteni	-347.0074	-93.6413
SO2	588.6096	1515.0233



Obr. 2: Počet úmrtí v Londýně od 1. do 15. 12. 1952 v závislosti na znečištění vzduchu a množství oxidu siřičitého přežívajících osob s diagnostikovaným nádorovým onemocněním kůže v populaci mužů kraje Praha

3 Čtyři „stejné“ regresní modely

Následující tabulka obsahuje čtyři dvojice naměřených hodnot (X_1, Y_1), (X_2, Y_2), (X_3, Y_3) a (X_4, Y_4). Spočtete odhady parametrů regresní přímky tato měření.

X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Pro jednotlivé modely dostáváme pomocí software R následující výstupy:

Model 1	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.0001	1.1247	2.67	0.0257
X1	0.5001	0.1179	4.24	0.0022

Model 2	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.0009	1.1253	2.67	0.0258
X2	0.5000	0.1180	4.24	0.0022

Model 3	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.0025	1.1245	2.67	0.0256
X3	0.4997	0.1179	4.24	0.0022

Model 4	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.0017	1.1239	2.67	0.0256
X4	0.4999	0.1178	4.24	0.0022

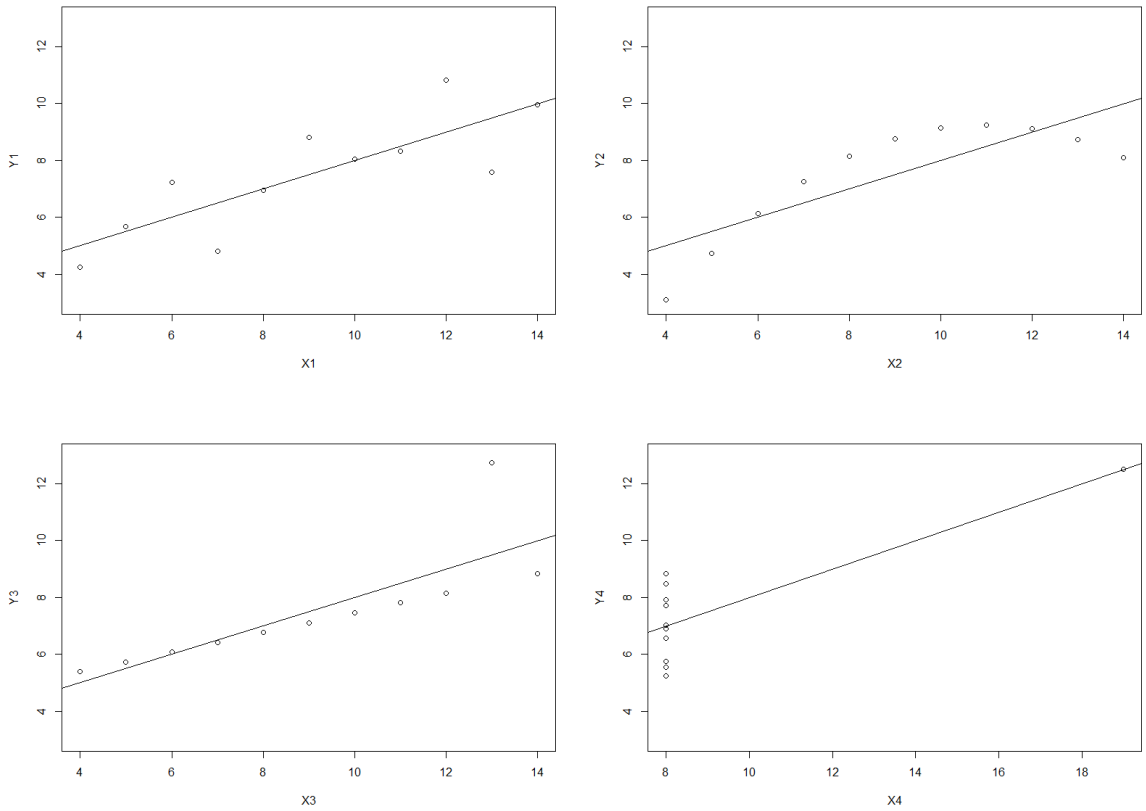
Výsledky regresních analýz uvedených v tabulkách ukazují, že je možné pro naprosto odlišné dvojice naměřených hodnot (X_1, Y_1), (X_2, Y_2), (X_3, Y_3) a (X_4, Y_4) získat identické odhady regresních funkcí i se stejnými hodnotami směrodatných chyb odhadnutých parametrů. Uvedené regresní modely jsou znázorněny na obrázku 3.

Příklady k procvičení

1. V tabulce jsou uvedeny počty výjezdů HZS Litomyšl (vysvětlovaná proměnná) od roku 2003 do roku 2012 (vysvětlující proměnná je čas). Cílem je popsat dynamiku těchto výjezdů pomocí vhodného regresního modelu. Ověřte vhodnost a adekvátnost daného modelu.

Rok	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Výjezdy HZS	540	594	643	681	784	830	905	1045	1232	1452

[Datový soubor: vyjezdyHZS.txt]



Obr. 3: Grafické zobrazení čtyř „stejných“ regresních modelů

2. Následující tabulka zachycuje údaje o počtech osob, které zahynuly při dopravních nehodách v letech 1994–2008 jak v České republice tak i v celé Evropské unii. V obou případech modelujte trend pomocí regresní přímky. Pomocí testu obecné lineární hypotézy odpovězte na otázku, zda je možné považovat obě odhadnuté regresní přímky za rovnoběžné.

Rok	1994	1995	1996	1997	1998	1999	2000	2001
CR	1637	1588	1562	1597	1360	1455	1486	1334
EU	63903	63155	59409	60267	58982	57691	56427	54302
Rok	2002	2003	2004	2005	2006	2007	2008	
CR	1431	1447	1382	1286	1063	1221	1076	
EU	53342	50351	47290	45346	43104	42496	38875	

[Datový soubor:: nehody.txt]

3. Následující tabulka uvádí hmotnost a systolický tlak 26 náhodně vybraných mužů ve věku od 25 do 30 let.

Hmotnost	Tlak	Hmotnost	Tlak
83	130	86	153
84	133	80	128
90	150	84	132
78	128	87	149
106	151	92	158
88	146	108	150
95	150	98	163
105	140	90	156
100	148	72	124
75	125	120	170
79	133	118	165
85	135	96	160
85	150	94	159

Pro daný datový soubor odhadněte parametry těchto modelů:
$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$
$Y_i = \beta_0 + \frac{\beta_1}{x_i} + \epsilon_i, i = 1, \ldots, n.$
Pro tyto modely určete reziduální rozptyl, index (koeficient) determinace, zkonstruujte 95% intervaly spolehlivosti pro parametry dané regresní přímky, proveďte testy významnosti regresních koeficientů a pomocí F-testu ověřte významnost modelu ($\alpha = 0{,}05$). Na základě předcházejících výpočtů zvolte vhodnější model – uveďte zdůvodnění vaší volby. Jaký systolický tlak lze očekávat u muže s hmotností 85 kg? Vypočtěte v tomto bodě 95% interval spolehlivosti pro regresní funkci a individuální předpověď. [Datový soubor: hmotnost_tlak.txt]

4. Tabulka uvádí část výsledků průzkumu spokojenosti v jedné nemocnici.

Věk	Závažnost	Stres	Spokojenost	Věk	Závažnost	Stres	Spokojenost
55	50	2,1	68	24	34	3,1	102
46	24	2,8	77	42	30	3	88
30	46	3,3	96	50	48	4,2	70
35	48	4,5	80	58	61	4,6	52
59	58	2	43	60	71	5,3	43
61	60	5,1	44	62	62	7,2	46
74	65	5,5	26	68	38	7,8	56
38	42	3,2	88	70	41	7	59
27	42	3,1	75	79	66	6,2	26
51	50	2,4	57	63	31	4,1	52
53	38	2,2	56	39	42	3,5	83
41	30	2,1	88	49	40	2,1	75
37	31	1,9	88				

Sestavte regresní model pro vysvětlovanou proměnnou popisující spokojenost pacientů (čím větší spokojenost, tím větší hodnota dané proměnné) v závislosti ne věku pacienta, závažnosti onemocnění (čím větší hodnota proměnné, tím je onemocnění závažnější) a indexem stresu (velké hodnoty tohoto indexu identifikují velkou stresovou zátěž). Posuďte přesnost odhadů jednotlivých regresorů, jejich významnost a významnost celého regresního modelu. [Datový soubor: nemocnice_pruzkum.txt]