Regresní analýza

Statistika II

Jiří Neubauer

Katedra ekonometrie FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Regresní analýza

Cíl regresní analýzy:

- stanovení formy (trendu, tvaru, průběhu) této závislosti pomocí vhodné funkce
- vystihnout pomocí regresní funkce průběh (trend) závislosti mezi X a Y na základě znalosti dvojic empirických hodnot $[x_i, y_i]$, kde $i = 1, 2, \ldots, n$.

Regresní přímka

Princip regresní analýzy nejdříve vysvětlíme na jednoduchém modelu dvou náhodných veličin X a Y, kde Y bude vysvětlovaná proměnná a X bude vysvětlující proměnná (**regresor**). Budeme předpokládat, že mezi vysvětlovanou proměnnou Y a vysvětlující proměnnou X platí přibližně lineární vztah. Měření nebo pozorování veličiny Y může být zatíženo náhodnou chybou e.

$$Y = \beta_1 + \beta_2 X + e,$$

kde β_1, β_2 jsou neznámé parametry (neznámé reálné konstanty), Y a e jsou náhodné veličiny a X je daná reálná proměnná. Dále předpokládáme, že při hodnotách $x_1, x_2, \ldots, x_n$ proměnné X pozorujeme hodnoty $y_1, \ldots, y_n$ proměnné Y zatížené chybami $e_1, \ldots, e_n$. Pozorování vyhovují modelu

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, \ n.$$

Regresní přímka

O chybách $e_1, \ldots, e_n$ předpokládáme, že jsou to nezávislé náhodné veličiny, že jsou **nesystematické**, tj. střední hodnota $E(e_i) = 0$, a **homogenní**, tj. že mají stejný rozptyl $D(e_i) = \sigma^2$, $i = 1, \ldots, n$. Cílem je najít odhad parametrů β_1, β_2 a σ^2. Použijeme k tomu **metodu nejmenších čtverců**. Označíme

$$S^2(\beta_1, \beta_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_1 + \beta_2 x_i))^2$$

součet čtverců náhodných chyb e_i a odhady $\widehat{\beta}_1, \widehat{\beta}_2$ parametrů β_1, β_2 stanovíme tak, aby součet čtverců chyb $S^2(\beta_1, \beta_2)$ nabyl minimální možné hodnoty.

Regresní přímka

Z matematiky je známo, že nutnou podmínkou pro existenci extrému funkce dvou a více proměnných je nulovost prvních parciálních derivací, tj. v našem případě

$$\frac{\partial S^2(\beta_1, \beta_2)}{\partial \beta_1} = \frac{\partial S^2(\beta_1, \beta_2)}{\partial \beta_2} = 0,$$

podmínku postačující pro minimum nemusíme vyšetřovat, neboť funkce $S(\beta_1, \beta_2)$ je ryze konvexní. Dostáváme tedy

$$\frac{\partial S^2(\beta_1, \beta_2)}{\partial \beta_1} = 2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)(-1) = 0,$$

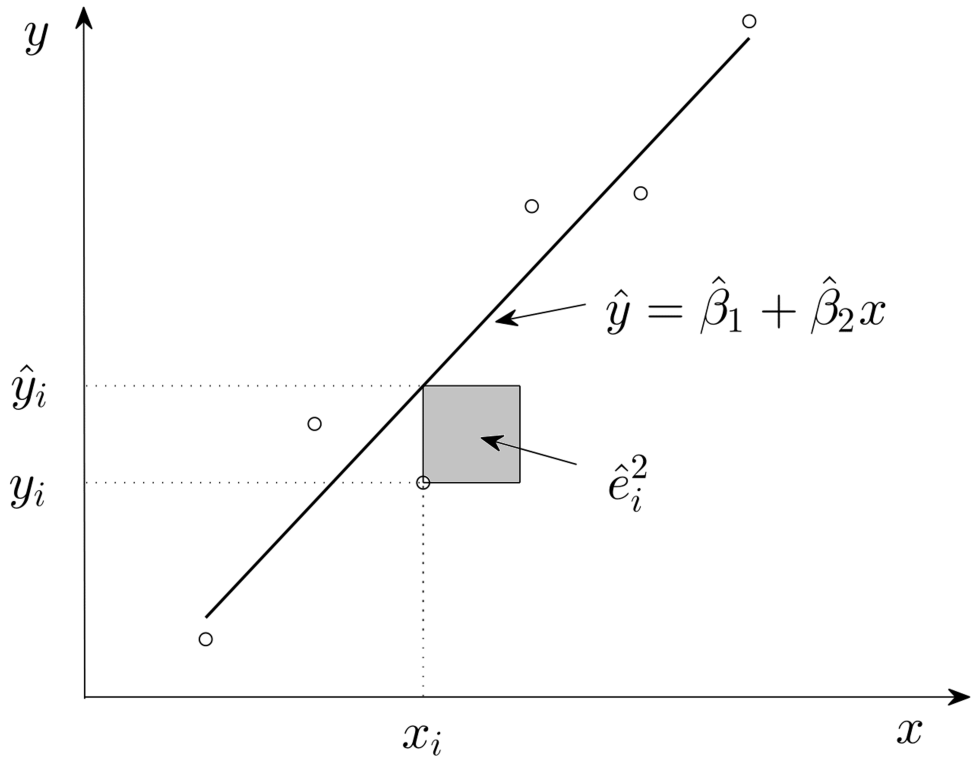
$$\frac{\partial S^2(\beta_1, \beta_2)}{\partial \beta_2} = 2 \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)(-x_i) = 0.$$

odkud získáme tzv. **soustavu normálních rovnic**

$$\beta_1 n + \beta_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

$$\beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Regresní přímka



Obrázek: Lineární regresní model – přímka

Regresní přímka

Vyřešíme-li tuto soustavu (např. Cramerovým pravidlem), obdržíme odhady parametrů

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}, \quad \widehat{\beta}_2 = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}.$$

Tyto odhady lze také vyjádřit ve tvaru

$$\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \overline{x} = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}, \quad \widehat{\beta}_2 = \frac{s_{xy}}{s_x^2},$$

kde $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ a $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$ jsou výběrové průměry,
$s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ je výběrový rozptyl a $s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$ je výběrová kovariance.

Regresní přímka

Přímku o rovnici $y = \beta_1 + \beta_2 x$ nazýváme **regresní přímkou**, β_1, β_2 jsou tzv. **regresní parametry (koeficienty)** a přímku o rovnici $\widehat{y} = \widehat{\beta}_1 + \widehat{\beta}_2 x$ nazýváme regresní přímkou s odhadnutými parametry $\widehat{\beta}_1$ a $\widehat{\beta}_2$. Hodnota $\widehat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i$ je predikovaná hodnota y v bodě x_i a veličiny $\widehat{e}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i$ nazýváme **rezidua**. Dále platí, že minimální hodnota součtu čtverců $S^2(\beta_1, \beta_2)$ je rovna

$$S_e = S^2\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \widehat{\beta}_1 \sum_{i=1}^{n} y_i - \widehat{\beta}_2 \sum_{i=1}^{n} x_i y_i.$$

S_e nazýváme **reziduální součet čtverců**. Je možné ukázat, že veličina $s_e^2 = \frac{1}{n-2} S_e$ je nevychýleným odhadem rozptylu σ^2, a tedy platí $E(s_e^2) = \sigma^2$.

Regresní přímka – příklad

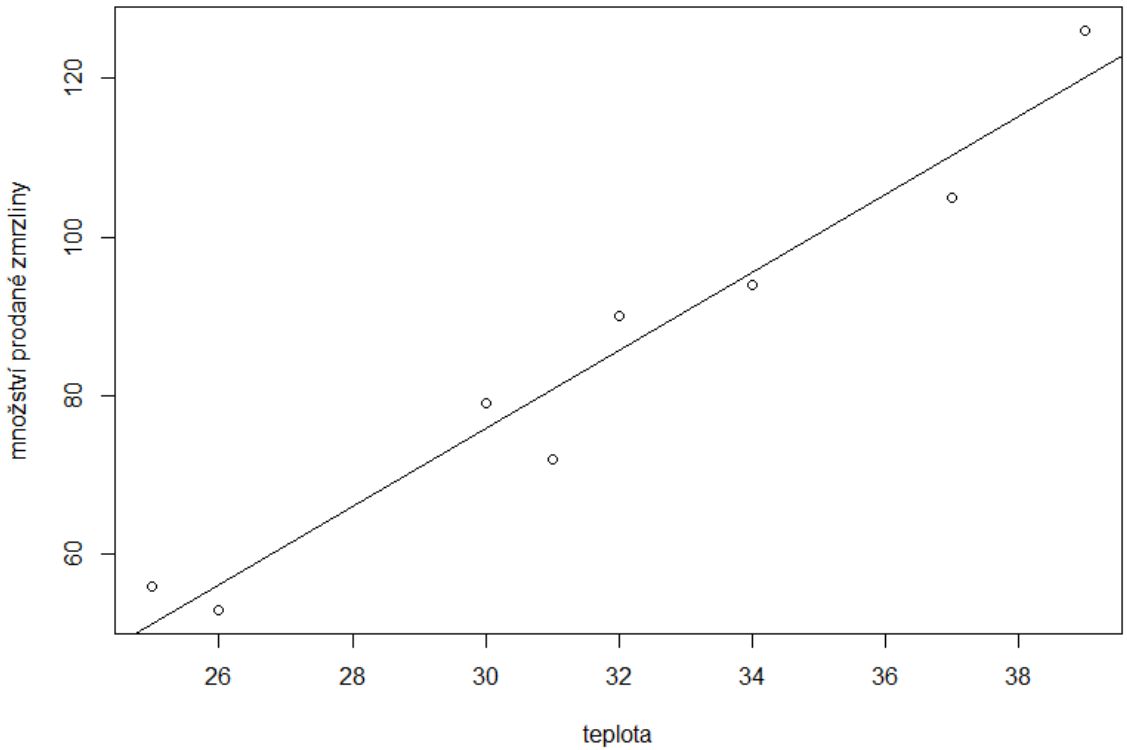
Následující tabulka udává informaci o teplotě (ve stupních Celsia) v jednom městě a množství zmrzliny (v kilogramech) prodaných v osmi náhodně vybraných cukrárnách.

teplota	34	30	25	32	37	39	31	26
zmrzlina	94	79	56	90	105	126	72	53

Vysvětlovanou proměnnou je v tomto případě množství zmrzliny, vysvětlující proměnnou potom teplota ve městě. Metodou nejmenších čtverců odhadneme parametry regresní přímky

$$\widehat{y} = -71{,}769 + 4{,}918x.$$

Regresní přímka – příklad



Obrázek: Regresní přímka – závislost množství prodané zmrzliny na teplotě

Lineární regresní model

Zobecníme předchozí výsledky a budeme předpokládat, že je potřeba modelovat nějakou sledovanou (hůře dostupnou či nesnadno měřitelnou) náhodnou veličinu Y (tzv. **vysvětlovaná veličina** nebo **odezva**) pomocí jiných snáze dostupných veličin $X_1, X_2, \ldots, X_k$ (nazývaných **vysvětlující proměnné** nebo **regresory**). Vyjdeme ze situace, kdy příslušná statistická data obsahují n nezávislých pozorování vysvětlované proměnné Y a odpovídajících n pozorování každého z regresorů $X_1, X_2, \ldots, X_k$. Budeme předpokládat, že i-té pozorování vysvětlované proměnné Y lze modelovat rovnicí:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i, \tag{1}$$

kde

1. y_i je i-té pozorování Y, $i = 1, \ldots, n$,
2. x_{ij} je i-té pozorování regresoru X_j, $i = 1, \ldots, n$, $j = 1, \ldots, k$,
3. β_j, $j = 1, \ldots, k$, jsou neznámé parametry,
4. e_i, $i = 1, \ldots, n$, jsou neznámé náhodné chyby, které vznikají při pozorování vysvětlované proměnné Y a které nemůžeme přímo pozorovat ani měřit.

Lineární regresní model

Přitom dále předpokládáme, že x_{ij} jsou pevně dané známé reálné hodnoty a veličiny Y_i a e_i jsou náhodného charakteru (náhodné veličiny). Na jejich pravděpodobnostní rozdělení klademe následující předpoklady:

(P1) Střední hodnota $E(e_i) = 0$, $i = 1, \ldots, n$, tj. náhodné chyby jsou **nesystematické**.

(P2) Rozptyl $D(e_i) = \sigma^2$, $i = 1, \ldots, n$, tj. náhodné chyby jsou **homogenní** se stejným neznámým rozptylem σ^2.

(P3) Náhodné chyby e_i jsou nezávislé.

Model daný rovnicí (1) spolu s předpoklady (P1), (P2), (P3) se nazývá **lineární regresní model** (LRM). Často se v lineárním regresním modelu předpokládá, že první regresor je konstanta, potom pozorované hodnoty $x_{i1} = 1, i = 1, \ldots, n$ a model má tvar

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i.$$

Funkci, která popisuje závislost vysvětlované proměnné Y na regresorech $X_1, X_2, \ldots, X_k$ pak nazýváme **regresní funkcí**.

Lineární regresní model

Odhad parametrů v lineárním regresním modelu (1) provedeme opět metodou nejmenších čtverců. Model nejdříve zapíšeme v maticovém tvaru. Označme:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Pak model (1) lze vyjádřit jednoduchým zápisem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Odhad neznámých parametrů pak stanovíme řešením soustavy lineárních rovnic

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \text{– tzv. } \textbf{normální rovnice}.$$

Lineární regresní model

Jejich řešení snadno nalezneme za předpokladu, že matice $\mathbf{X}'\mathbf{X}$ je regulární a tedy existuje inverzní matice $(\mathbf{X}'\mathbf{X})^{-1}$. Za tohoto předpokladu říkáme, že model je plné hodnosti. V modelu plné hodnosti lze řešení normálních rovnic zapsat ve tvaru

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Pro reziduální součet čtverců zapsaný v maticovém tvaru pak dostaneme vyjádření

$$S_e = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}.$$

Dále budeme pracovat jenom s modely plné hodnosti.

Lineární regresní model – regresní parabola

Uvedeme nyní dva příklady lineárních regresních modelů: regresní paraboly a modelu se dvěma lineárními regresory. Nejprve budeme uvažovat model, kdy vysvětlovaná proměnná Y je kvadratickou funkcí vysvětlující proměnné X, tvaru:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + e_i, \quad i = 1, \ldots, n.$$

Zřejmě jde o speciální případ LRM (lineárního vzhledem k neznámým parametrům β_1, β_2, β_3). V maticovém zápisu tohoto modelu je:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^2 y_i \end{pmatrix}.$$

Lineární regresní model – regresní parabola

Za předpokladu, že model je plné hodnosti, lze odhad $\widehat{\boldsymbol{\beta}}$ vektoru $\boldsymbol{\beta}$ získat řešením rovnic $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ ve tvaru $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Potom lze reziduální součet čtverců S_e vyjádřit ve tvaru

$$S_e = \sum_{i=1}^{n} y_i - \widehat{\beta}_1 \sum_{i=1}^{n} y_i - \widehat{\beta}_2 \sum_{i=1}^{n} x_i y_i - \widehat{\beta}_3 \sum_{i=1}^{n} x_i^2 y_i$$

a odhad rozptylu σ^2 je $s_e^2 = S_e/(n-3)$.

Lineární regresní model – regresní parabola

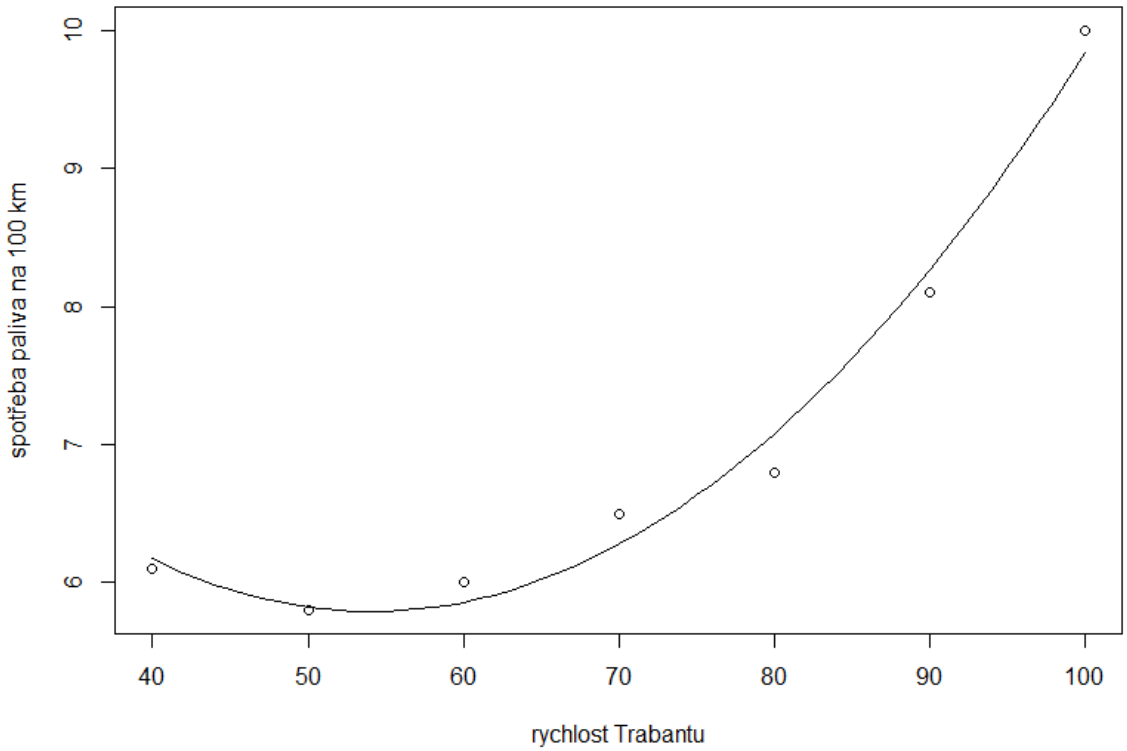
Příklad. U automobilu Trabant se měřila spotřeba paliva v litrech na 100 km (Y) v závislosti na jeho rychlosti (X).

Rychlost	40	50	60	70	80	90	100
Spotřeba	6,1	5,8	6,0	6,5	6,8	8,1	10,0

Odhadnutá parabolická regresní funkce má tvar

$$\widehat{y} = 11{,}39386 - 0{,}20726x + 0{,}001917x^2.$$

Lineární regresní model – regresní parabola



Obrázek: Regresní parabola – závislost spotřeby paliva na rychlosti

Lineární regresní model – dva lineární regresory

Předpokládejme, že vysvětlovaná proměnná Y může záviset na dvou regresorech X a Z (používáme označení X místo X_1 a Z místo X_2, které je v aplikacích tohoto typu časté). K dispozici je n nezávislých pozorování veličiny Y při daných n hodnotách veličin X a Z. Vyjdeme z modelu

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i, \quad i = 1, \ldots, \ n,$$

který je speciálním případem obecného lineárního regresního modelu $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

Lineární regresní model – dva lineární regresory

Matice v modelu mají tvar

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix}, \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n z_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n z_i & \sum_{i=1}^n x_i z_i & \sum_{i=1}^n z_i^2 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n z_i y_i \end{pmatrix}.$$

Pak užitím metody nejmenších čtverců dostaneme odhad $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Lineární regresní model – dva lineární regresory

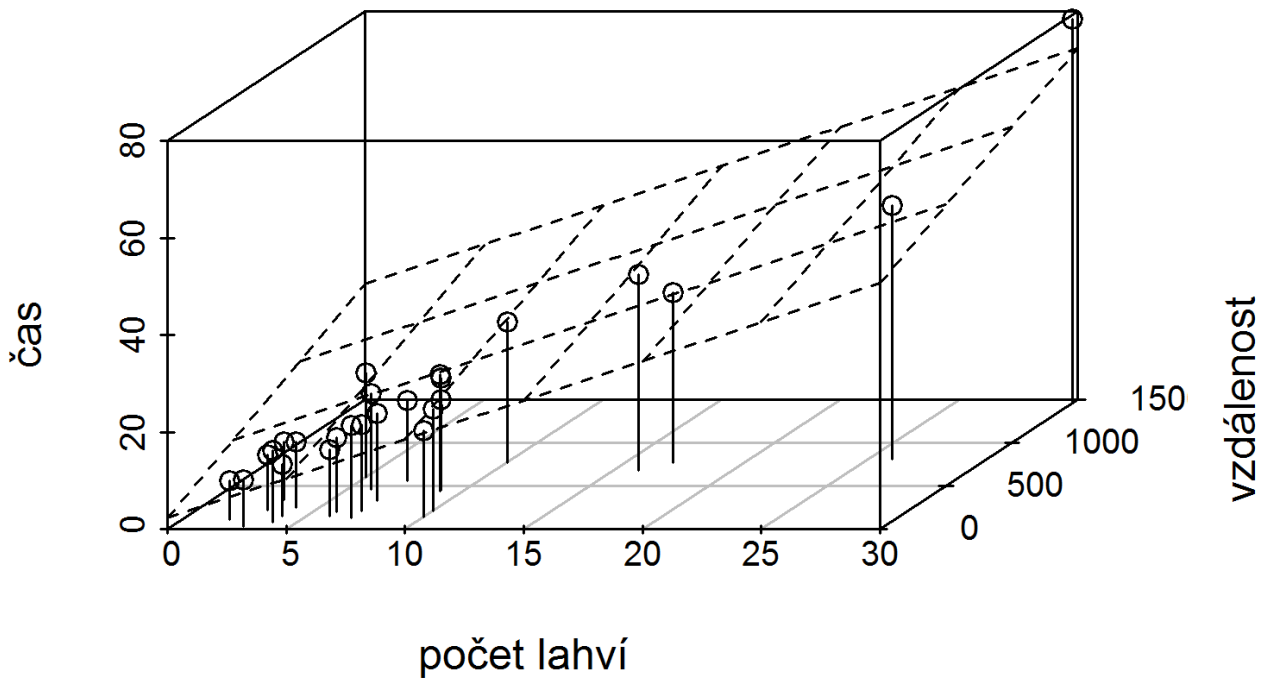
Příklad. Výrobce nealkoholických nápojů má zájem analyzovat potřebný čas k servisu (doplnění lahví případně malý servis zařízení) automatů na výdej lahví s těmito nápoji. Celkovou dobu doplnění lahví je třeba predikovat pomocí dvou dostupných proměnných: počet lahví, které je třeba doplnit do automatu, a vzdálenost, kterou musí údržbář ujít. Vysvětlovanou proměnnou je v tomto případě celkový čas, vysvětlující proměnné jsou počet doplněných lahví a vzdálenost.

čas	16,68	11,5	12,03	14,88	13,75	18,11	8	17,83	79,24	21,5
počet lahví	7	3	3	4	6	7	2	7	30	5
vzdálenost	560	220	340	80	150	330	110	210	1460	605
čas	40,33	21	13,5	19,75	24	29	15,35	19	9,5	35,1
počet lahví	16	10	4	6	9	10	6	7	3	17
vzdálenost	688	215	255	462	448	776	200	132	36	770
čas	17,9	52,32	18,75	19,83	10,75					
počet lahví	10	26	9	8	4					
vzdálenost	140	810	450	635	150					

Lineární regresní model – dva lineární regresory

Metodou nejmenších čtverců získáme odhad regresní funkce

$$\widehat{y} = 2,341 + 1,616x + 0,014z.$$



Obrázek: Regrese se dvěma lineárními regresory – závislost času potřebného na servis na počtu případů doplňování automatu a vzdálenosti, kterou musí údržbář ujít

Volba regresní funkce

Některé typy lineárních regresních funkcí:

- přímková regrese $Y = \beta_1 + \beta_2 X$,
- hyperbolická regrese $Y = \beta_1 + \frac{\beta_2}{X}$,
- logaritmická regrese $Y = \beta_1 + \beta_2 \ln X$,
- parabolická regrese $Y = \beta_1 + \beta_2 X + \beta_3 X^2$
- polynomická regrese $Y = \beta_1 + \beta_2 X + \cdots + \beta_p X^p$

Některé typy nelineárních regresních funkcí:

- exponenciální regrese $Y = \beta_1 \beta_2^X$,
- mocninná regrese $Y = \beta_1 X^{\beta_2}$.