11. Regresní analýza

Čas ke studiu kapitoly: 60 minut

Cíl Po prostudování tohoto odstavce budete umět

- vysvětlit pojem obecný lineární model
- princip lineárního regresního modelu
- používat výsledky regresní analýzy
- verifikovat regresní model pomocí indexu determinace

VÝKLAD

11.1. Úvod

Regrese

Pod pojmem regrese rozumíme systematické změny jedněch veličin při změnách jiných veličin a popis těchto změn matematickými funkcemi. Snažíme se tedy napozorované hodnoty vyrovnat vhodnou matematickou funkcí. Celá výstavba regresního modelu bude mít několik fází. Jedná se především o

- předběžnou analýzu dat (výpočet základních charakteristik, grafický průběh, studium věcných vztahů mezi veličinami apod.)
- výběr vhodné funkce, zahrnující
  - odhad modelu - volba vhodného postupu při odhadu parametrů regresní funkce
  - verifikace modelu

Závislost jevů a veličin

- funkční závislost veličiny Y na veličině X ve tvaru $y=f(x)$, kde hodnotám proměnné X jsou jednoznačně přiřazeny hodnoty Y
- pravděpodobnostní pojetí - z teorie pravděpodobnosti vyplývá, že dva jevy považujeme za závislé, jestliže nastoupení kteréhokoliv z nich ovlivňuje pravděpodobnost nastoupení druhého jevu
- statistická závislost - systematický pohyb hodnot jedné veličiny při růstu či poklesu hodnot druhé veličiny. Jde přitom o stochastický vztah mezi těmito veličinami.

Terminologie

Vysvětlovaná (závisle) proměnná - proměnná v regresním modelu, jejíž chování se snažíme vysvětlit, popsat matematickou křivkou. Tato proměnná vystupuje v modelu jako výsledek působení tzv. vysvětlujících proměnných. Jedná se tedy o proměnnou na levé straně regresní funkce a většinou ji označujeme symbolem *Y*.

Vysvětlující (nezávisle) proměnné - proměnné v regresním modelu, jejichž chování vysvětluje chování závisle proměnné *Y*. Tyto proměnné vystupují v modelu jako příčinné proměnné, to znamená, že v důsledku jejich změny se mění vysvětlovaná proměnná. Jedná se tedy o proměnné na pravé straně regresní funkce a většinou je označujeme symbolem *X*, *Z* apod.

Poznámka: Pojem levá a pravá strana regresní rovnice je samozřejmě relativní, jde spíše o zažitou konvenci, která se však důsledně dodržuje. Totéž se týká i používaného značení.

11.2. Obecný lineární model

Celá regresní analýza je založena na obecnějším pojmu, zvaném lineární model. Obecným lineárním modelem rozumíme model ve tvaru (maticový zápis)

$$\underline{Y} = X\,\underline{\beta} + \underline{e}$$

kde

- $\underline{Y}$ je náhodný vektor n hodnot vysvětlované proměnné
- $\boldsymbol{X}$ je matice zadaných hodnot vysvětlujících proměnných o rozměrech $n \times (k)$
- $\underline{\beta}$ je vektor p neznámých parametrů ($p=k$)
- $\underline{e}$ je vektor n hodnot náhodných chyb

Předpoklady obecného lineárního modelu

1. $E(e_i) = 0$ pro každé $i=1,2,\ldots,n$
   Střední hodnota náhodné složky je nulová. Tato podmínka znamená, že náhodná složka nepůsobí systematickým způsobem na hodnoty vysvětlované proměnné *Y*.
2. $D(e_i) = \sigma^2$ pro každé $i=1,2,\ldots,n$
   Rozptyl náhodné složky je konstantní (hovoříme o tzv. *homoskedasticitě*). Tato podmínka vyjadřuje, že variabilita náhodné složky nezávisí na hodnotách vysvětlujících proměnných a tudíž i podmíněná variabilita vysvětlované proměnné nezávisí na hodnotách vysvětlujících proměnných a je rovna neznámé kladné konstantě σ^2.
3. Cov $(e_i, e_j)=0$ pro každé $i \neq j$, kde $i, j = 1,2,\ldots,n$
   Kovariance náhodné složky je nulová. Tedy hodnoty náhodné složky jsou nekorelované a z toho vyplývá i nekorelovanost různých dvojic pozorování vysvětlované proměnné *Y*.
4. $\boldsymbol{X}$ je nestochastická (nenáhodná) matice. Znamená to tedy, že vysvětlující proměnné jsou nenáhodné.
5. Parametry β_j, $j=1,2,\ldots,k$ mohou nabývat libovolných hodnot. Na vektor $\boldsymbol{\beta}$ tedy nejsou kladeny žádné omezující podmínky.

Pokud budou platit ještě další předpoklady 6 a 7, pak tento lineární model se nazývá **regresní**:

6. Matice $\boldsymbol{X}$ má plnou hodnost, tedy $h(\boldsymbol{X})=k$ a dále $n > k$ (n je počet pozorování).
   Tato podmínka vyžaduje, aby mezi vysvětlujícími proměnnými nebyla funkční lineární závislost, tedy v matici $\boldsymbol{X}$ nesmí existovat lineárně závislé sloupce. Počet

vysvětlujících proměnných nesmí být pochopitelně větší než počet pozorování a v praxi by být měl počet pozorování výrazně větší než počet vysvětlujících proměnných.

7. e_i mají normální rozdělení pravděpodobnosti pro každé $i=1,2,\ldots,n$.
   Z této podmínky vyplývá normalita i pro vysvětlovanou proměnnou Y. Náhodný vektor $\underline{Y}$ má potom n-rozměrné normální rozdělení s vektorem středních hodnot $\boldsymbol{X}\underline{\beta}$ a kovarianční maticí $\sigma^2 I_n$.

11.3. Základní regresní modely

- **Obecná regresní přímka,** nebo **lineární regrese s jednou vysvětlující proměnnou** (nejpoužívanější): $Y_i = \beta_0 + \beta_1 . x_i + e_i$

  pro speciální matici $X = \begin{pmatrix} 1 & x_1 \\ . & . \\ 1 & x_n \end{pmatrix}$; $\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

- **Kvadratická regrese**: $Y_i = \beta_0 + \beta_1 . x_i + \beta_2 . x_i^2 + e_i$

  pro speciální matici $X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ . & . & . \\ 1 & x_n & x_n^2 \end{pmatrix}$; $\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$

- **Regrese se dvěma nezávislými proměnnými**: $Y_i = \beta_0 + \beta_1 . x_i + \beta_2 . z_i + e_i$

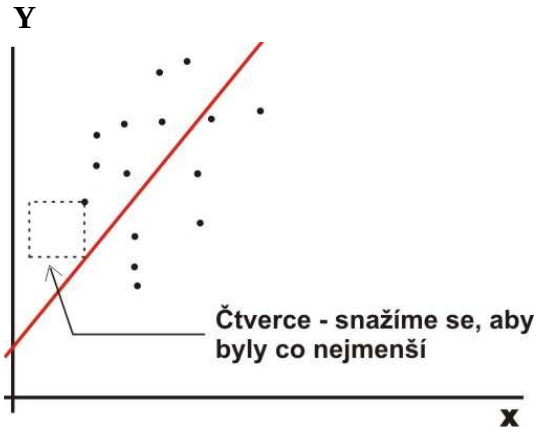
  pro speciální matici $X = \begin{pmatrix} 1 & x_1 & z_1 \\ . & . & . \\ 1 & x_n & z_n \end{pmatrix}$; $\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$

- **Nelineární model**: $Y_i = f(x_i, z_i, \underline{\beta})$ ... řeší se většinou tak, že se převádí na model lineární. Např.
  $Y_i = \beta_0 \beta_1^{x_i} \beta_2^{z_i}$, který lze přepsat do lineárního tvaru (lineárního v parametrech)
  $\ln Y_i = \ln(\beta_0) + x_i \ln(\beta_1) + z_i \ln(\beta_2)$

11.4. Lineární regrese s jednou vysvětlující proměnnou

Mějme $n > 2$ pozorování, tedy n dvojic $(Y_i, x_i);\ i=1,..,n$

$Y_i = \beta_0 + \beta_1 . x_i + \boldsymbol{e_i}$



Určení modelu: pomocí **metody nejmenších čtverců**, tj. z podmínky, aby výraz

$$\varphi = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot x_1)^2 = \sum_{i=1}^{n} (e_i)^2$$ byl minimální.

Nalezení koeficientů:

$$\begin{cases} min\,\varphi \Rightarrow \beta_0, \beta_1 (odhady\ \ parametrů\ označ.\ b_0, b_1) \\ \dfrac{\partial \varphi}{\partial \beta_0} = 0; \dfrac{\partial \varphi}{\partial \beta_1} = 0 \end{cases},$$

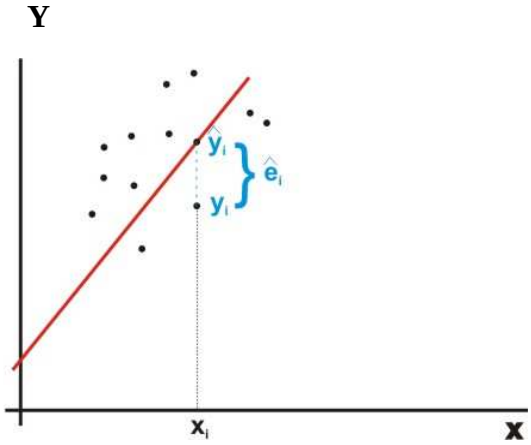
což je tzv. soustava normálních rovnic, kterou vyřešíme:

$$b_0 = \overline{Y} - b_1 \cdot \overline{x} = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \cdot Y_i$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot Y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Odhadem očekávané hodnoty E(Y| x) regresní funkce pro libovolné x je statistika

$$\hat{Y}(x) = b_0 + b_1 x = \ldots = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \overline{x}) \cdot (x - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \cdot Y_i$$



$\hat{e}_i = Y_i - \hat{Y}(x_i) = Y_i - \hat{Y}_i$ … chyba mezi skutečnou a modelovou hodnotou tzv. **reziduum.**

Položíme dále

$SS_R = \sum_{i=1}^{n} (\hat{e}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ … **reziduální součet čtverců**

$S_R^2 = \frac{SS_R}{n-2}$ …**reziduální rozptyl.**

Dá se ukázat, že následující statistika

$\frac{S_R^2}{\sigma^2} \cdot (n-2) \sim \chi^2 (n-2)$ tj. má rozdělení chí-kvadrát s *(n-2)* stupni volnosti.

Střední hodnoty a rozptyly získaných odhadů b_0, b_1:

1. $Eb_0 = \beta_0$; $Eb_1 = \beta_1$; $E\hat{Y}(x) = \beta_0 + \beta_1 x$;

2. $$Db_0 = \sigma_{b_0}^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)^2 = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{(n-1)s_x^2} \right)$$; kde $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

dále podobně

$$Db_1 = \sigma_{b_1}^2 = \frac{\sigma^2}{(n-1)s_x^2}$$

a konečně

$$D\left[\hat{Y}(x)\right] = \sigma_{\hat{Y}}^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})\cdot(x - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)^2 = \ldots = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right)$$

Pomocí následujících statistik provedeme odhady těchto rozptylů:
položíme

$$S_{b_0}^2 = S_R^2 \cdot \frac{\sigma_{b_0}^2}{\sigma^2} \; ; \; S_{b_1}^2 = S_R^2 \cdot \frac{\sigma_{b_1}^2}{\sigma^2} ; \; S_{\hat{Y}}^2 = S_R^2 \cdot \frac{\sigma_{\hat{Y}}^2}{\sigma^2}$$

Snadno se přesvědčíme, že jsou to nestranné odhady příslušných rozptylů.

Testy hypotéz a intervaly spolehlivosti

Na základě předpokladu normality popisovaného regresního modelu lze usoudit, že

$$\frac{b_0 - \beta_0}{\sigma_{b_0}} \sim \mathrm{N}(0,1) \; ; \; \frac{b_1 - \beta_1}{\sigma_{b_1}} \sim \mathrm{N}(0,1) \; ; \frac{\hat{Y}(x) - \beta_0 - \beta_1 x}{\sigma_{\hat{Y}}} \sim \mathrm{N}(0,1)$$

A na základě statistického chování reziduálního rozptylu víme, že

$$\frac{b_0 - \beta_0}{S_{b_0}} \sim t(n\text{-}2) \; ; \; \frac{b_1 - \beta_1}{S_{b_1}} \sim t(n\text{-}2) \; ; \; \frac{\hat{Y}(x) - \beta_0 - \beta_1 x}{S_{\hat{Y}}} \sim t(n\text{-}2) \; ;$$

tj. všechny uvedené výběrové statistiky mají Studentovo rozdělení s (n-2) stupni volnosti. Toho lze samozřejmě využít jak pro účely testování hypotéz, tak i pro konstrukci intervalových odhadů.

Dílčí t-testy

Dílčí t-testy jsou testy o hodnotách jednotlivých parametrů regresní funkce a umožňují nám testovat oprávněnost setrvání vysvětlující proměnné v regresním modelu. Testujeme (postupně pro jednotlivá i) nulovou hypotézu ve tvaru

H_0: $\beta_i = 0$ *pro* $i=0,1$
proti alternativě
H_A: $\beta_i \neq 0$ *pro* $i=0,1$

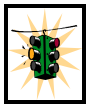
Pokud se ukáže, že pro konkrétní *i* nelze zamítnout nulovou hypotézu, je třeba zvážit setrvání příslušné vysvětlující proměnné v modelu. Pokud by se totiž parametr u příslušné proměnné neodlišoval významně od nuly, pak taková proměnná do modelu nic nového nepřináší a je v něm tudíž zbytečně. „Nadbytečnost“ proměnné v modelu by se však měla prokázat i podle jiných kritérií. Dále je však třeba poznamenat, že z hlediska kvality výsledných odhadů prováděných na základě regresního modelu je horší variantou případ, kdy proměnnou, která do modelu patří, chybně vyřadíme (testování hypotéz - chyba II. druhu) než případ, kdy proměnná do modelu nepatří a my ji tam chybně ponecháme (chyba I. druhu). Přitom je třeba si uvědomit, že pod kontrolou máme pouze pravděpodobnost chyby I. druhu, nikoliv však již pravděpodobnost chyby II. druhu.

Závěrem je třeba poznamenat, že vyřazení (či nové zařazení) proměnné do modelu znamená spustit celý proces tvorby modelu od začátku a tedy znamená to i nový odhad regresních parametrů.

Testové statistiky pro výše uvedené dílčí nulové hypotézy jsou odvozené Studentovy t-statistiky s *n-2* stupni volnosti:

$$\frac{b_0 - \beta_0}{S_{b_0}} \sim t(n-2) \;;\; \frac{b_1 - \beta_1}{S_{b_1}} \sim t(n-2) \;;$$

Současná výpočetní technika a především statistické pakety, jako např. STATGRAPHIC nám umožňují číst výsledky takovýchto testů přímo v podobě výstupních hodnot p-value, jak demonstruje dále vyřešený příklad.



Řešený příklad

Firma provádí opravy stolních kalkulátorů a pokladen. Data zapsána v tabulce pocházejí z 18 ohlášených oprav. U každé opravy je uveden počet opravovaných kalkulátorů x a celková doba opravy (v minutách) Y.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
x	7	6	5	1	5	4	7	3	4	2	8	5	2	5	7	1	4	5
Y	97	86	78	10	75	62	101	39	53	33	118	65	25	71	105	17	49	68

a) Nalezněte odhady koeficientů regresní přímky.
b) Zakreslete data a regresní funkci.
c) Proveďte dílčí t-testy o hodnotách jednotlivých parametrů regresní funkce.

Řešení

– pomocí programového paketu STATGRAPHIC:

Lineární regrese - Doba opravy vs. Počet

Regression Analysis - Linear model: Y = b_0 + b_1*x

Dependent variable: Doba opravy
Independent variable: Počet

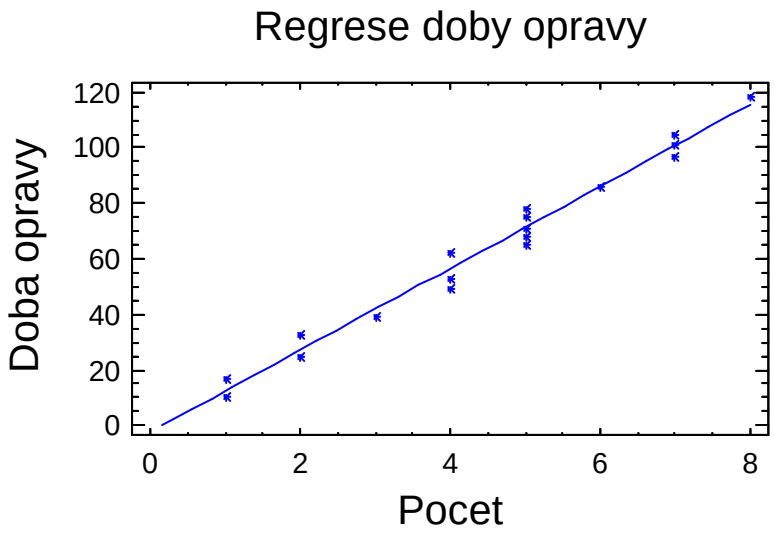
Parameter	Estimate	Standard Error	T Statistic	P-Value
b_0 - Intercept	-2,32215	2,56435	-0,905549	0,3786
b_1 - Slope	14,7383	0,519257	28,3834	0,0000

V kontextu s předchozími odvozenými vztahy je nyní označeno:
$b_0 = Intercept$, $b_1 = Slope$, obě hodnoty uvedeny ve druhém sloupci. Ve třetím sloupci jsou pak uvedeny pozorované hodnoty S_{b_0}, S_{b_1}.

Následující funkce představuje rovnici pro odhad očekávané hodnoty doby opravy:

Doba opravy = -2,32215 + 14,7383 . Počet

Pozorované hodnoty testových statistik pro dílčí t-testy jsou uvedeny v předposledním sloupci (T Statistic), příslušné hodnoty p-value jsou pak v posledním sloupci. Z výsledku je patrné, že hypotézu H_0: $\beta_0=0$ nezamítneme s ohledem na významnou hodnotu v příslušném sloupci p-value. Na základě toho můžeme prohlásit, že regresní přímka prochází počátkem, což je i logický závěr s ohledem na povahu dat. Druhý z dílčích testů nám říká, že směrnice přímky (Slope) je hodnota, která se významně liší od nuly, neboť jsme zamítli hypotézu H_0: $\beta_1=0$.



Interval spolehlivosti pro očekávanou hodnotu E(Y | x)

Bodovým odhadem očekávané hodnoty Y pro zadanou hodnotu x , tedy E(Y| x) = $\beta_0 + \beta_1 x$, je statistika

$$\hat{Y}(x) = b_0 + b_1 x = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x}) \cdot (x - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \cdot Y_i$$

Při hledání intervalového odhadu pro E(Y| x) budeme vycházet zejména z výše odvozené *t*-statistiky:

$$\frac{\hat{Y}(x) - \beta_0 - \beta_1 x}{S_{\hat{Y}}} \sim t(n\text{-}2) .$$

Z ní a na základě běžného postupu, aplikovaného při hledání intervalového odhadu, můžeme získat snadno následující intervalový odhad pro E(Y| x), se spolehlivostí α:

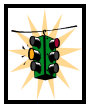
$$\text{E(Y| x)} = \beta_0 + \beta_1 x \in \left\langle \hat{Y}(x) - S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2), \hat{Y}(x) + S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2) \right\rangle$$

Tyto intervalové meze pro spojitě se měnící hodnoty x tvoří tzv. **pás spolehlivosti kolem regresní přímky.** Šířka tohoto pásu je závislá na hodnotě $S_{\hat{Y}}$. V některých aplikacích se můžeme setkat s otázkou, pro kterou volbu x je pás spolehlivosti nejužší, a tudíž také interval spolehlivosti pro očekávanou hodnotu E(Y| x) nejpřesnější ? Tento problém lze zodpovědět nalezením takového x_{opt}, které minimalizuje $S_{\hat{Y}}$:

$$S_{\hat{Y}}^2 = S_R^2 \cdot \frac{\sigma_{\hat{Y}}^2}{\sigma^2} = S_R^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right)$$

$$\Rightarrow \quad x_{opt} = \bar{x}$$

Vidíme, že pás má nejmenší šířku pro $x_{opt} = \bar{x}$, a při změně *x* ať už k větším či menším hodnotám šířka pásu monotónně roste. Šířku pásu lze do určité míry předem ovlivnit vhodnou volbou bodů $(x_1, \ldots, x_n)$.

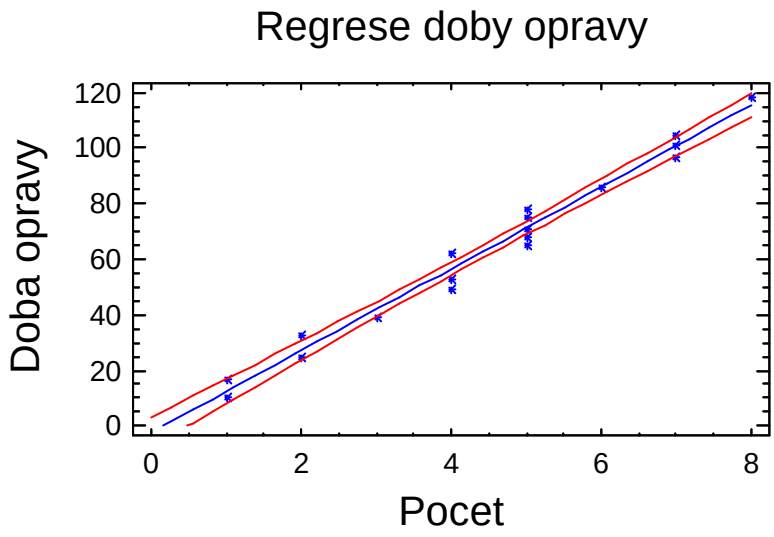


Řešený příklad – pokračování

d) Nalezněte 95% pás spolehlivosti kolem regresní přímky pro dobu opravy v závislosti na počtu kalkulátorů

e) Nalezněte bodový a intervalový odhad pro očekávanou dobu opravy pěti kalkulátorů.

Řešení



Pro $x=5$ dostáváme:

$$\hat{Y}(x) = b_0 + b_1 x = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x}) \cdot (x - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \cdot Y_i = 71.3691$$

$$\mathrm{E}(Y|\, x) = \beta_0 + \beta_1 x \in < \hat{Y}(x) - S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2), \hat{Y}(x) + S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2) >= < 69.063,\ 73.6752 >$$

Index determinace

Slouží pro účely verifikace správnosti zvoleného regresního modelu. Při aplikaci metody nejmenších čtverců platí vztah $SS_Y = SS_{\hat{Y}} + SS_R$,

kde $SS_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ je celkový součet čtverců,

$SS_{\hat{Y}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ je součet čtverců modelu a

$SS_R = \sum_{i=1}^{n} (\hat{e}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ je reziduální součet čtverců.

U součtu čtverců modelu by se ve vzorci místo průměru z napozorovaných hodnot měl spíše objevit průměr z hodnot odhadnutých. Při aplikaci metody nejmenších čtverců se však dá odvodit, že tyto průměry jsou stejné, lze tedy psát

$$\bar{Y} = \bar{\hat{Y}}$$

Je zřejmé, že čím je model lepší, tím větších hodnot bude nabývat součet čtverců modelu a reziduální součet čtverců bude menší. Naopak špatný model znamená velkou hodnotu reziduálního součtu čtverců ve srovnání se součtem čtverců modelu. Celou rovnost můžeme vydělit celkovým součtem čtverců a převést tak na tvar

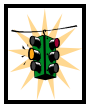
$$1 = \frac{SS_{\hat{Y}}}{SS_Y} + \frac{SS_R}{SS_Y}$$

Oba zlomky jsou kladné, jejich součet je roven jedničce, tedy nutně musí být hodnota obou zlomků mezi nulou a jedničkou. Pro příslušné zlomky platí nyní analogická úvaha jako pro samotné součty čtverců. Bude-li model dobře vystihovat závislost vysvětlované proměnné na pravé straně rovnice (tedy na vysvětlujících proměnných), poroste hodnota prvního zlomku v rovnosti k jedničce a druhý zlomek se bude blížit k nule. Bude-li model popisovat uvažovanou závislost špatně, bude tomu naopak. Je tedy logické vzít první zlomek jako kritérium kvality regresního modelu. Položíme tedy

$$I^2 = \frac{SS_{\hat{Y}}}{SS_Y}$$

a nazveme jej **indexem determinace**. Index determinace tedy

- udává kvalitu regresního modelu, přesněji řečeno udává, kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno;
- nabývá hodnot od nuly do jedné (teoreticky i včetně těchto krajních mezí), přičemž hodnoty blízké nule značí špatnou kvalitu regresního modelu; hodnoty blízké jedné značí dobrou kvalitu regresního modelu;
- udává se většinou v procentech.



Řešený příklad – pokračování

f) Posuďte kvalitu vyšetřeného modelu lineární regrese pro dobu opravy v závislosti na počtu kalkulátorů pomocí indexu determinace.

Řešení

$$SS_Y = SS_{\hat{Y}} + SS_R$$

Zdroj		Součty čtverců
Model	$SS_{\hat{Y}}$	16182,6
Residuální	SS_R	321,396
Celkový	SS_Y	16504,0

$$I^2 = \frac{SS_{\hat{Y}}}{SS_Y} = 98.0526\ \%$$

Poznámka

V případě lineární regrese s více vysvětlujícími proměnnými má však index determinace jednu nepříjemnou vlastnost, která částečně snižuje jeho kvalitu. Závisí totiž na počtu vysvětlujících proměnných a s růstem jejich počtu narůstá i jeho hodnota. Proto se častěji než

samotný index determinace používá tzv. modifikovaný index determinace, který je „penalizovaný“ za nadbytečný počet vysvětlujících proměnných. Má tvar

$$I_M^2 = I^2 - \frac{(1-I^2)(p-1)}{n-p}$$

kde p je počet odhadovaných parametrů v modelu. Jeho hodnota je tedy vždy nepatrně menší než hodnota indexu nemodifikovaného.

Shrnutí pojmů

Regresní model je speciální případ **obecného lineárního modelu.** Základními předpoklady jsou nulovost střední hodnoty chyb, dále pak homoskedasticita a předpoklad normality rozdělení chyb.

Vysvětlovaná (závisle) proměnná je proměnná v regresním modelu, která je náhodná a jejíž chování se snažíme vysvětlit, popsat matematickou křivkou.
Vysvětlující (nezávisle) proměnné jsou proměnné v regresním modelu, jejichž chování vysvětluje chování závisle proměnné.
Lineární regresní model s jednou vysvětlující proměnnou je základním modelem a je založen na **metodě nejmenších čtverců.** Z ní lze odvodit parametry tohoto modelu s velmi příznivými statistickými vlastnostmi. Součet čtverců odchylek skutečných od modelových hodnot se nazývá **reziduální součet čtverců**.

Dílčí t-testy jsou testy o hodnotách jednotlivých parametrů regresní funkce a umožňují nám testovat oprávněnost setrvání vysvětlující proměnné v regresním modelu.
Na základě příznivých statistických vlastností odhadovaných parametrů modelu můžeme získat snadno intervalový odhad pro očekávanou hodnotu vysvětlované proměnné $E(Y|x)$, se spolehlivostí α. Tyto intervalové meze pro spojitě se měnící hodnoty x tvoří tzv. **pás spolehlivosti kolem regresní přímky.** Šířka tohoto pásu je nejmenší pro výběrový průměr: $x_{opt} = \bar{x}$.
Index determinace slouží pro účely verifikace správnosti zvoleného regresního modelu

Úlohy k řešení

Př. 1:

Při kontrolních měřeních rozměrů silikátových štítových dílců bylo náhodně vybráno 8 dílců vykazujících vesměs kladné odchylky v délce i výšce od normovaných hodnot:

odchylka délky [mm]	3	4	4	5	8	10	6	3
odchylka výšky [mm]	4	6	5	6	7	13	9	4

Najděte lineární regresní model závislosti odchylky výšky na odchylce délky.

Př. 2:

V letech 1931-1961 byly měřeny průtoky v profilu nádrže Šance na Ostravici a v profilu nádrže Morávka na Morávce. Roční průměry v m^3/s jsou dány v následující tabulce:

rok	Šance	Morávka
1931	4,130	2,476
1932	2,386	1,352
1933	2,576	1,238
1934	2,466	1,725
1935	3,576	1,820
1936	2,822	1,913
1937	3,863	2,354
1938	3,706	2,268
1939	3,710	2,534
1940	4,049	2,308
1941	4,466	2,517
1942	2,584	1,726
1943	2,318	1,631
1944	3,721	2,028
1945	3,290	2,423

rok	Šance	Morávka
1946	2,608	1,374
1947	2,045	1,194
1948	3,543	1,799
1949	4,055	2,402
1950	2,224	1,019
1951	2,740	1,552
1952	3,792	1,929
1953	3,087	1,488
1954	1,677	0,803
1955	2,862	1,878
1956	3,802	1,241
1957	2,509	1,165
1958	3,656	1,872
1959	2,447	1,381
1960	2,717	1,679

Předpokládejte, že v jednom z následujících let chybí hodnota průměrného ročního průtoku pro nádrž Morávka. V tomto roce činil průměrný roční průtok v profilu nádrže Šance na Ostravici 2,910 m^3/s. Na základě lineární regrese odhadněte hodnotu průměrného ročního průtoku nádrže Morávka.