14 JEDNODUCHÁ REGRESE

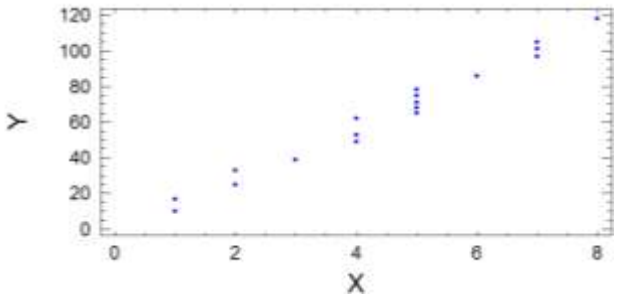
Čas ke studiu kapitoly: 120 minut

Cíl: Po prostudování této kapitoly budete

- rozumět základním pojmům regresní analýzy
- znát zjednodušující předpoklady regresního modelu
- umět používat metodu nejmenších čtverců pro odhad regresní funkce
- umět odhadnout důvěryhodnost odhadnuté regresní funkce pomocí pásu spolehlivosti pro $E(Y \mid X=x_0)$ a pásu predikce
- umět posoudit vhodnost modelu pomocí indexu determinace
- umět používat interpolaci a extrapolaci a budete si vědomi rizik s tím spojených

Výklad:

V praxi většinou nestudujeme náhodné veličiny jako takové, zajímá nás jejich vztah k jiným náhodným veličinám. Vysoký stupeň závislosti (korelace) často odráží příčinný vztah, ale nemusí tomu tak být vždy. **Příčinné souvislosti** (**kauzalitu**) čistě empirickými prostředky neodhalíme. Ke statistickým výsledkům je třeba přidat odborné znalosti a praktické zkušenosti. V nejjednodušším případě je souvislost mezi sledovanými znaky zcela jednoznačná. Například hmotnost předmětů, které jsou homogenní, je funkci jejich objemu. Závislost tohoto druhu se nazývá **funkční závislost**. Předmětem statistiky je však hodnocení



takových závislostí, kdy neexistuje zcela jednoznačný vztah mezi sledovanými znaky. Tento vztah označujeme jako **regresi**. Při měření závislosti dvou kvantitativních znaků můžeme druh a sílu závislosti orientačně posoudit z bodového grafu (**korelačního pole**), v němž je každá dvojice údajů graficky znázorněna jedním bodem v rovině. Druh závislosti odhadujeme pomocí křivky, která se dobře hodí k napozorovaným hodnotám. Podle typu křivky rozeznáváme závislost lineární, logaritmickou, exponenciální a další. Jedním z úkolů regresní analýzy dat je i vyjádření síly závislosti mezi sledovanými znaky, tj. stanovení, do jaké míry je hodnota jednoho znaku předurčena hodnotou druhého znaku. V této kapitole se se budeme zabývat nejjednodušším případem, kdy zkoumáme závislost jedné proměnné (Y) na jedné proměnné (X) a tato závislost je lineární.

Řešený příklad:

Pro snazší pochopení problematiky uvažujme konkrétní případ: 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.

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

Vyneseme-li si do grafu závislost celkové doby opravy (Y) na počtu opravovaných kalkulátorů (X), získáme následující bodový graf označovaný také jako **korelační pole**:

Z grafu se zdá být zřejmé, že počet opravovaných kalkulátorů ovlivňuje celkovou dobu opravy. Naučíme se, jak toto popsat pomocí vyrovnávací křivky, jak používat vyrovnávací křivku k prognózám a jak vyhodnotit vhodnost volby typu vyrovnávací křivky.

Výklad:

14.1 Pojmy

Nejdříve se seznámíme se základní terminologii.

Vysvětlovaná (závisle) proměnná - proměnná v regresním modelu, jejíž chování se snažíme vysvětlit, popsat **vyrovnávací 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. (V našem případě jde o celkovou dobu opravy.)

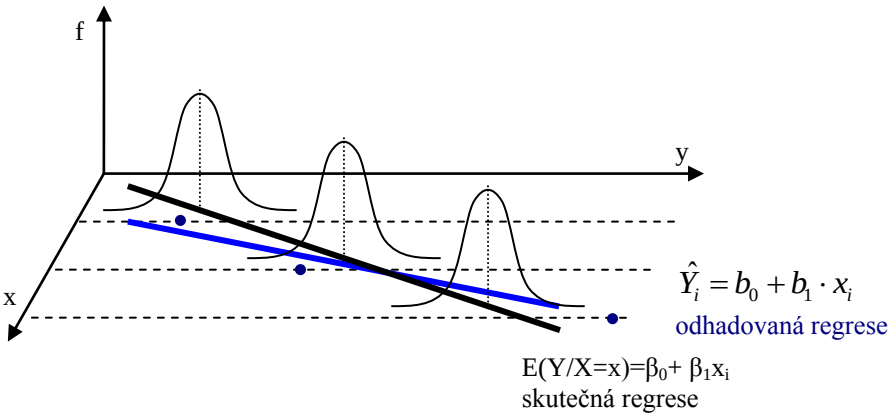
Vysvětlující (nezávisle) proměnná - proměnná v regresním modelu, jejíž chování vysvětluje chování závisle proměnné Y. Tato proměnná vystupuje v modelu jako příčinná proměnná, to znamená, že v důsledku její změny se mění vysvětlovaná proměnná. Jedná se tedy o proměnnou na pravé straně regresní funkce a většinou je označujeme symbolem X. (V našem případě jde o počet opravovaných kalkulátoru.)

***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.*

Reziduum (chyba predikce) $e_i = \left(Y_i - \hat{Y}_i\right)$ – odchylka hodnoty předpovídané vyrovnávací křivkou $\left(\hat{Y}_i\right)$ a skutečně naměřené hodnoty $\left(Y_i\right)$.

Regresní funkce – $EY_i = \beta_0 + \beta_1 \cdot x_i$, skutečná regrese populace, v praxi je neznámá a musíme ji odhadovat na základě pozorovaní $\left[x_i, Y_i\right]$. Odhad regrese má tvar: $\hat{Y}_i = b_0 + b_1 \cdot x_i$

Vraťme se k našemu příkladu. Dokázali byste od oka proložit bodovým grafem vyrovnávací přímku? Nakolik by byla tato přímka vyhovující? V případě, kdy jsou body grafu značně rozptýleny musíme použít objektivnější metodu než „od oka“. V následující části se budeme zabývat metodou algebraických výpočtů pro nalezení vyrovnávací křivky.

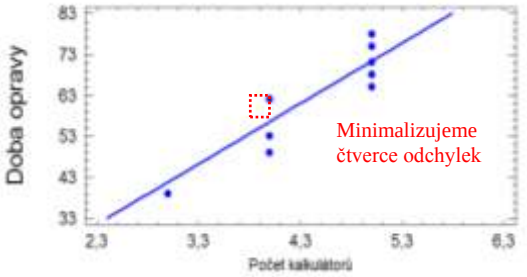


14.2 Metoda nejmenších čtverců

Naším cílem je najít vyrovnávací přímku, jejíž rovnice má tvar:

$$\hat{Y}_i = b_0 + b_1 \cdot x_i$$

b_0 a b_1 musíme zvolit tak, abychom získali co nejméně rozptýlený soubor vertikálních odchylek $e_i = \left(Y_i - \hat{Y}_i\right)$, tzv. **chyb predikce, resp. reziduí**.



Nejdříve nás napadne, že bychom mohli minimalizovat $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)$. Avšak některé body se nacházejí pod přímkou, jiné nad přímkou, proto by některé odchylky byly kladné, jiné záporné, vzájemně by se rušily … Abychom se tomu vyhnuli, mohli bychom minimalizovat součet jejich absolutních odchylek. Vzhledem k tomu, že minimalizace funkce se provádí pomocí její derivace (vzpomeňte si na derivaci „absolutní hodnoty“), není ani toto vhodná metoda. Mnohem známější a tudíž i mnohem používanější je tzv. **metoda nejmenších čtverců**, která spočívá v minimalizaci součtů kvadrátů reziduí.

Mějme alespoň 2 pozorování (n>2) o souřadnicích $[x_i;Y_i]$.
Součet čtverců reziduí:

$$\varphi = \sum_{i=1}^{n}\left(Y_i - b_0 - b_1 x_i\right)^2$$

Součet čtverců reziduí minimalizujeme:

$$\frac{d\varphi}{db_0} = -2\sum_{(i)}\left(Y_i - b_0 - b_1 \cdot x_i\right) = 0$$
$$\frac{d\varphi}{db_1} = -2\sum_{(i)}\left[\left(Y_i - b_0 - b_1 \cdot x_i\right)\cdot\left(x_i\right)\right] = 0$$

Danou soustavu upravíme na tvar:

$$\sum_{(i)} Y_i - nb_0 - b_1\sum_{(i)} x_i = 0$$
$$\sum_{(i)} x_i Y_i - b_0\sum_{(i)} x_i - b_1\sum_{(t)} x_i^2 = 0$$

Řešení nalezneme ve tvaru:

$$b_0 = \frac{\sum_{(i)} Y_i}{n} - b_1\frac{\sum_{(i)} x_i}{n} = \bar{Y} - b_1 \cdot \bar{x}$$

$$b_1 = \frac{n\sum_{(i)} x_i Y_i - \sum_{(i)} x_i \sum_{(i)} Y_i}{n\sum_{(i)} x_i^2 - \left(\sum_{(i)} x_i\right)^2}$$

Vztahy pro výpočet koeficientů b_0 a b_1 odvodíme v jednodušší podobě – v tzv. **odchylkové formě**, věnujeme-li nyní trochu času vhodnějšímu vyjádření $\hat{Y}_i$.

$$\hat{Y}_i = b_0 + b_1 x_i = \left(b_0 + b_1 \cdot \bar{x}\right) + b_1 \cdot \left(x_i - \bar{x}\right) = b_0^* + b_1 \cdot \left(x_i - \bar{x}\right)$$

Součet čtverců reziduí :

$$\varphi = \sum_{(i)} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{(i)} \left(Y_i - b_0^* - b_1 \cdot \left(x_i - \bar{x}\right)\right)^2$$

Součet čtverců reziduí minimalizujeme:

$$\frac{d\varphi}{db_0} = -2\sum_{(i)} \left(Y_i - b_0^* - b_1 \cdot \left(x_i - \bar{x}\right)\right) = 0$$
$$\frac{d\varphi}{db_1} = -2\sum_{(i)} \left[\left(Y_i - b_0^* - b_1 \cdot \left(x_i - \bar{x}\right)\right) \cdot \left(x_i - \bar{x}\right)\right] = 0$$

Danou soustavu upravíme na tvar:

$$\sum_{(i)} Y_i - nb_0^* - b_1 \sum_{(i)} \left(x_i - \bar{x}\right) = 0$$
$$\sum_{(i)} \left(x_i - \bar{x}\right) \cdot Y_i - b_0^* \sum_{(i)} \left(x_i - \bar{x}\right) - b_1 \sum_{(i)} \left(x_i - \bar{x}\right)^2 = 0$$

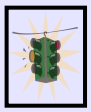
Řešení nalezneme ve tvaru:

$$b_0^* = \frac{\sum_{(i)} Y_i}{n} - b_1 \frac{\sum_{(i)} \left(x_i - \bar{x}\right)}{n} = \bar{Y} \qquad \Rightarrow \qquad b_0 = b_0^* - b_1 \cdot \bar{x} = \bar{Y} - b_1 \cdot \bar{x}$$

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

Poznámka: Využili jsme toho, že $\sum_{i=1}^{n} \left(x_i - \bar{x}\right) = 0$.

Vyrovnávací přímka má tedy tvar: $\hat{Y}_i = b_0 + b_1 x_i = \overline{Y} - b_1 \cdot \overline{x} + b_1 x_i = \overline{Y} + b_1\left(x_i - \overline{x}\right)$, z čehož je zřejmé, že **vždy prochází bodem** $\left[\overline{x}, \overline{Y}\right]$.



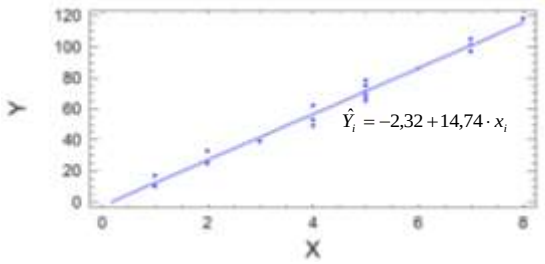
Řešený příklad:

Výpočet koeficientů vyrovnávací přímky v našem případě:

x_i	7	6	5	1	5	4	7	3	4	2	8	5	2	5	7	1	4	5	$\overline{x} = 4{,}5$
Y_i	97	86	78	10	75	62	101	39	53	33	118	65	25	71	105	17	49	68	$\overline{Y} = 64{,}0$
$\left(x_i - \overline{x}\right)$	2,50	1,50	0,50	-3,50	0,50	-0,50	2,50	-1,50	-0,50	-2,50	3,50	0,50	-2,50	0,50	2,50	-3,50	-0,50	0,50	$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$
$\left(x_i - \overline{x}\right)^2$	6,25	2,25	0,25	12,25	0,25	0,25	6,25	2,25	0,25	6,25	12,25	0,25	6,25	0,25	6,25	12,25	0,25	0,25	$\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = 74{,}5$
$\left(x_i - \overline{x}\right)\cdot Y_i$	242,5	129,0	39,0	-35,0	37,5	-31,0	252,5	-58,5	-26,5	-82,5	413,0	32,5	-62,5	35,5	262,5	-59,5	-24,5	34,0	$\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot Y_i = 1098{,}0$

$$b_1 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot Y_i}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \frac{1098{,}0}{74{,}5} = 14{,}74\,, \qquad b_0 = \overline{Y} - b_1 \cdot \overline{x} = 64{,}0 - 14{,}74 \cdot 4{,}5 = -2{,}32$$

$$\Rightarrow \underline{\underline{\hat{Y}_i = -2{,}32 + 14{,}74 \cdot x_i}}$$



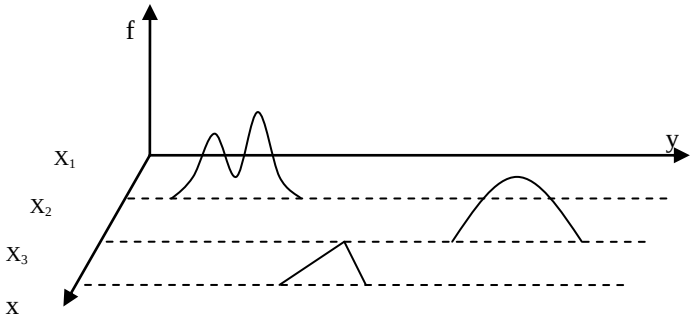
Výklad:

Až dosud byl náš přístup k výběru popisný. Data jsme pouze nahradili vyrovnávací přímkou. Nyní potřebujeme učinit úsudky o populaci, z níž výběr pochází. Za tím účelem potřebujeme sestrojit statistický model, který nám umožní sestrojit intervaly spolehlivosti a testovat hypotézy.

14.3 Regresní model

Předpokládejme že určitý počet kalkulátorů x_1 jsme přidělili několika pracovníkům. Celková doba opravy nebude u všech stejná. Někteří pracovníci mají větší zkušenosti, někteří měli smůlu a byly jim přiděleny kalkulátory s komplikovaným odstraněním poruchy, apod. Takto

vytvoříme populaci hodnot Y, správněji řečeno rozdělení pravděpodobnosti Y_1 na úrovni x_1 $f(Y_1|x_1)$. Podobně můžeme sestrojit také rozdělení $f(Y_2|x_2)$ atd. Pak můžeme znázornit množinu rozdělení Y takto:



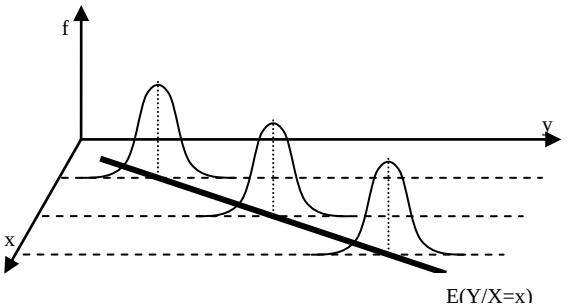
Analýza takovýchto rozdělení by byla obtížná. Aby byl problém zvládnutelný, stanovíme si předpoklady ohledně rozdělení Y:

1. **Linearita:** Pro každé rozdělení Y_i platí, že střední hodnota $E(Y_i|X_i) = E(Y_i) = \mu_i$ leží na přímce o které víme, že je skutečnou regresní přímkou (regresní přímkou populace, $Y_i = \beta_0 + \beta_1 x_i$).

2. **Homogenní rozptyl:** Všechna Y_i mají stejný rozptyl.

3. **Nezávislost:** Náhodné veličiny Y_i jsou navzájem statisticky nezávisle.

4. **Normalita:** Náhodné veličiny Y_i mají pro $i = 1,2,\ldots,n$ normální rozdělení



V některých případech je vhodné využít při zápisu regresní přímky rezidua e_i, neboli odchylky Y_i od její střední hodnoty. Alternativní zápis regresního modelu pak vypadá takto:

$$Y_i = \beta_0 + \beta_1 \cdot x_i + e_i ,$$

kde

1. $E(e_i) = 0$ pro každé *i=1,2,...,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 (homoskedasticitní). 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. **Normalita:** Náhodné složky e_i mají pro $i = 1,2,\ldots,n$ normální rozdělení

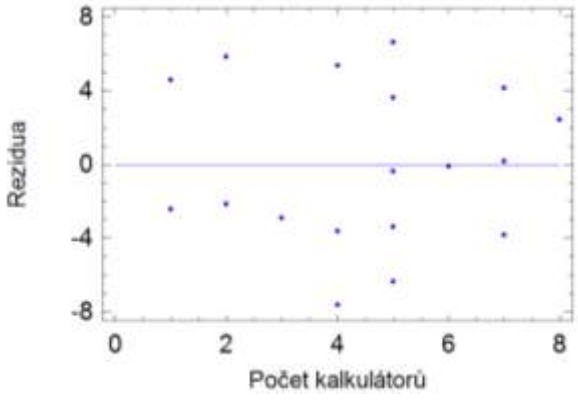
Proto, abychom mohli model nazvat lineárním regresním modelem, musí být splněny ještě následující dvě podmínky:

5. **Regresní parametry β_i mohou nabývat libovolných hodnot**.

6. **Regresní model je lineární v parametrech**.

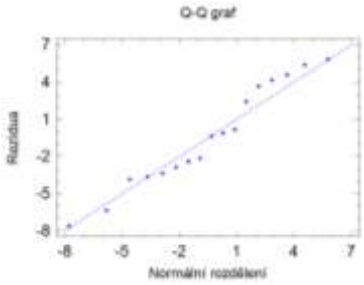
Předpoklady na nichž je model založen ověřujeme většinou pomocí jednoduchých exploratorních grafů, resp. pomocí známých testů .

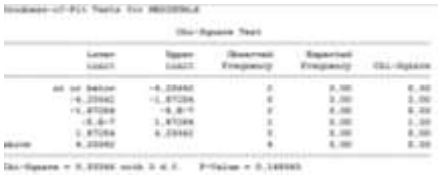
Porovnání reziduí s čímkoli dalším (pozorovanými hodnotami, odhadnutými hodnotami, hodnotami *X*) by nemělo ukázat žádné systematické závislosti. Nejužitečnější je v takovém případě často graf reziduí a předvídaných hodnot.



Rezidua jsou náhodně rozmístěna kolem nuly a nemají žádný zřejmý vztah k předpovídaným hodnotám: ani se systematicky nezvyšují ani se systematicky nesnižují spolu s rostoucími předpovídanými hodnotami a není zde ani náznak nelineárního vztahu.

Protože předpokládáme, že kolísání hodnot závisle proměnné kolem regresní přímky je dáno normálním rozdělením, rezidua by se měla chovat alespoň přibližně jako výběr z normálního rozdělení s nulovou střední hodnotou. Q-Q graf reziduí by tedy měl být přibližně přímkou. Normalitu a nulovou střední hodnotu reziduí můžeme ověřit například pomocí Chí-kvadrát testu dobré shody a t-testu střední hodnoty.





14.4 Odhady koeficientů regresní přímky (β_0 a β_1)

Pro nalezení intervalových odhadů β_0 a β_1 potřebujeme znát střední hodnoty a rozptyly $\hat{Y}_0$, b_0 a b_1.

14.4.1 Střední hodnota a rozptyl b_1

Jaký je význam koeficientu b_1? Podle definice udává koeficient b_1 směrnici (sklon) vyrovnávací přímky, což je změna Y v závislosti na změně x, tzn.:

b_1 udává změnu závisle proměnné Y při jednotkové změně nezávislé proměnné x.

Např. v našem motivačním případě je b_1 14,74, tzn., že zvýšíme-li pracovníkovi počet kalkulátoru o 1, pak se celková doba pro opravu kalkulátoru zvedne o 14,74 minut.

Jaké je rozdělení b_1 kolem hledané hodnoty β_1 nám dává informaci o tom, jak blízko je odhadovaná přímka skutečné regresní přímce populace.

1. pravidlo normální aproximace pro regresi

Odhad koeficientu b_1 je přibližně normálně rozdělen se střední hodnotou $E(b_1)=\beta_1$ a rozptylem $D(b_1)=\sigma_{b_1}^2=\frac{\sigma^2}{(n-1)\cdot s_x^2}$.

Vidíme, že existují tři způsoby, jak snížit rozptyl b_1:

1. Snížení σ (rozptýlenost Y_i, reziduální směrodatná odchylka)
2. Zvýšení n (rozsah výběru)
3. Zvýšení s_x (rozptýlenost x_i)

Zvýšení s_x nazýváme **protiváhou** hodnot x_i k b_1.

Průvodce studiem:

Tento průvodce je opět určen zájemcům o matematické pozadí použitých vztahů. Je věnován odvození střední hodnoty a rozptylu b_1.

Metodou nejmenších čtverců jsme odvodili, že $b_1 = \dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot Y_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$.

Napíšeme-li si tento výraz explicitně, odvodíme jednoduše střední hodnotu a rozptyl odhadu b_1.

$$b_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot Y_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\left(x_1 - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\cdot Y_1 + \frac{\left(x_2 - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\cdot Y_2 + \ldots + \frac{\left(x_n - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\cdot Y_n =$$

$$= w_1 \cdot Y_1 + w_2 \cdot Y_2 + \ldots + w_n \cdot Y_n$$

kde $w_i = \dfrac{\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad i = 1,2,...,n$

Střední hodnota b_1

Protože x_i *a tím i* w_i *jsou konstanty, platí:*

$$E(b_1) = w_1 \cdot EY_1 + w_2 \cdot EY_2 + \ldots + w_n \cdot EY_n =$$

$$= w_1 \cdot (\beta_0 + \beta_1 x_1) + w_2 \cdot (\beta_0 + \beta_1 x_2) + \ldots + w_n \cdot (\beta_0 + \beta_1 x_n) =$$

$$= \beta_0 \cdot (w_1 + w_2 + \ldots + w_n) + \beta_1 \cdot (w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n) = \beta_1$$

Poznámka:

Využili jsme toho, že:

$$(w_1 + w_2 + \ldots + w_n) = \frac{\left(x_1 - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} + \frac{\left(x_2 - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} + \ldots + \frac{\left(x_n - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} =$$

$$= \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = 0$$

$$\left(w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n\right) = \frac{\left(x_1 - \bar{x}\right) \cdot x_1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} + \frac{\left(x_2 - \bar{x}\right) \cdot x_2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} + \ldots + \frac{\left(x_n - \bar{x}\right) \cdot x_n}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} =$$

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

$$= \frac{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}{\sum_{i=1}^{n} x_i^2 - 2 \cdot n \cdot \bar{x}^2 + n \cdot \bar{x}^2} = \frac{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2} = 1$$

Rozptyl b_1

V našem regresním modelu předpokládáme, že Y_i jsou nezávislé, proto rozptyl jejich lineární kombinace můžeme jednoduše vyjádřit jako:

$$D(b_1) = w_1^2 \cdot DY_1 + w_2^2 \cdot DY_2 + \ldots + w_n^2 \cdot DY_n$$

Model rovněž předpokládá, že všechna Y_i mají stejný rozptyl σ^2, proto:

$$D(b_1) = \sigma_{b_1}^2 = w_1^2 \cdot \sigma^2 + w_2^2 \cdot \sigma^2 + \ldots + w_n^2 \cdot \sigma^2 = \sigma^2 \cdot \left(w_1^2 + w_2^2 + \ldots + w_n^2\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sigma^2}{(n-1) \cdot s_x^2}$$

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} = \frac{\sigma}{\sqrt{n-1} \cdot s_x}$$

neboť:

$$\left(w_1^2 + w_2^2 + \ldots + w_n^2\right) = \frac{\left(x_1 - \bar{x}\right)^2}{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^2} + \frac{\left(x_2 - \bar{x}\right)^2}{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^2} + \ldots + \frac{\left(x_n - \bar{x}\right)^2}{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^2} =$$

$$= \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^2} = \frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{1}{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} \cdot (n-1)} = \frac{1}{s_x^2 \cdot (n-1)}$$

Výklad:

14.4.2 Střední hodnota a rozptyl b_0

2. pravidlo normální aproximace pro regresi

Odhad koeficientu b_0 je přibližně normálně rozdělen se střední hodnotou $E(b_0)=\beta_0$ a rozptylem $Db_0=\sigma_{b_0}^2=\sigma^2\cdot\left(\frac{1}{n}-\frac{\bar{x}^2}{(n-1)\cdot s_x^2}\right)=\frac{\sum_{i=1}^{n}(Y_i-\hat{Y})^2}{n-2}\cdot\left(\frac{1}{n}-\frac{\bar{x}^2}{(n-1)\cdot s_x^2}\right)$.

Odvození:

Střední hodnota b_0

$$b_0=\bar{Y}-b_1\cdot\bar{x}$$
$$Eb_0=E\bar{Y}-E(b_1\cdot\bar{x})=\bar{Y}-\beta_1\cdot\bar{x}=\beta_0$$

Poznámka: Využili jsme toho, že regresní přímka prochází bodem $[\bar{x},\bar{Y}]$.

Rozptyl b_0

$$b_0=\bar{Y}-b_1\cdot\bar{x}$$

$$Db_0=\sigma_{b_0}^2=D\bar{Y}-D(b_1\cdot\bar{x})=0-Db_1\cdot\bar{x}^2=\frac{\sigma^2}{n}-\frac{\sigma^2}{(n-1)\cdot s_x^2}\cdot\bar{x}^2=\sigma^2\cdot\left(\frac{1}{n}-\frac{\bar{x}^2}{(n-1)\cdot s_x^2}\right)$$

$$=\sigma^2\cdot\left(\frac{1}{n}-\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

14.4.3 Střední hodnota a rozptyl $\hat{Y}_0$

3. pravidlo normální aproximace pro regresi

$\hat{Y}_0=\hat{Y}(x_0)$, tj. odhad koeficientu Y_0 je přibližně normálně rozdělen se střední hodnotou $E\hat{Y}_0=\beta_0+\beta_1\cdot x_0$ a rozptylem $D\hat{Y}_0=\sigma^2\cdot\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)\cdot s_x^2}\right)=\sigma^2\cdot\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$.

Pro odvození $E\hat{Y}_0$ a $D\hat{Y}_0$ je vhodné využít odchylkové formy vyrovnávací přímky. Tzn. budeme uvažovat vyrovnávací přímku ve tvaru:

$$\hat{Y}_0 = b_0^* + b_1 \cdot \left(x_0 - \bar{x}\right)$$

Střední hodnota $\hat{Y}_0$

$$E\hat{Y}_0 = E\left(b_0^* + b_1 \cdot \left(x_0 - \bar{x}\right)\right) = E\left(b_0^*\right) + \left(x_0 - \bar{x}\right)E\left(b_1\right) = E\bar{Y} + \beta_1 \cdot \left(x_0 - \bar{x}\right) = \left(\bar{Y} - \beta_1 \cdot \bar{x}\right) + \beta_1 \cdot x_0 =$$
$$= \beta_0 + \beta_1 \cdot x_0$$

Rozptyl $\hat{Y}_0$

$$D\hat{Y}_0 = D\left(b_0^* + b_1 \cdot \left(x_0 - \bar{x}\right)\right) = D\left(b_0^*\right) + \left(x_0 - \bar{x}\right)^2 D\left(b_1\right) = D\left(\bar{Y}\right) + \left(x_0 - \bar{x}\right)^2 \cdot \frac{\sigma^2}{\left(n-1\right)\cdot s_x^2} =$$

$$= \frac{\sigma^2}{n} + \left(x_0 - \bar{x}\right)^2 \cdot \frac{\sigma^2}{\left(n-1\right)\cdot s_x^2} = \sigma^2 \cdot \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\left(n-1\right)\cdot s_x^2}\right) = \sigma^2 \cdot \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)$$

14.4.4 Interval spolehlivosti a testy pro β_1

Zavedli jsme si pojmy normalita, střední hodnota a rozptyl b_1, můžeme tedy přistoupit k intervalovým odhadům β_1.

Víme, že směrodatná odchylka b_1 je $\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} = \frac{\sigma}{\sqrt{n\text{-}1}\cdot s_x}$, přičemž **σ** označuje směrodatnou odchylku pozorování Y_i kolem regresní přímky populace (tzv. **reziduální směrodatnou odchylku**). σ je však obecně neznámá, proto i ji musíme odhadovat. Odhadem σ je výběrová směrodatná odchylka Y_i kolem vyrovnávací přímky, přičemž vezmeme v úvahu 2 stupně volnosti[1]:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}{n-2}}$$

s nazýváme **výběrová reziduální směrodatná odchylka.**

S využitím tohoto odhadu můžeme říci, že:

[1] Pokud bychom měli 2 pozorování, vyrovnávací přímku jimi proložíme jednoznačně. Nezbývá nám však žádná informace o rozptylu pozorování kolem vyrovnávací přímky. Informaci o rozptylu získáme pouze tehdy, máme-li k dispozici více než 2 pozorování. Tzn. použijeme-li rozptyl kolem vyrovnávací přímky k odhadu rozptylu kolem regresní přímky, zbývá nám (n-2) stupňů volnosti.

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

Na základě předpokladu normality popisovaného regresního modelu lze usoudit, že

$$b_1 \to N\left(\beta_1; \sigma_{b_1}^2\right) \;\Rightarrow\; \frac{b_1 - \beta_1}{\sigma_{b_1}} \to N(0;1)$$

a na základě statistického chování reziduálního rozptylu víme, že

$$\frac{b_1 - \beta_1}{s_{b_1}} \to t_{n-2}$$

Pomocí této výběrové statistiky pak můžeme zkonstruovat **interval spolehlivosti pro β_1**:

$$P(t_{\frac{\alpha}{2},n-2} < \frac{b_1 - \beta_1}{s_{b_1}} < t_{1-\frac{\alpha}{2},n-2}) = 1-\alpha$$

$$P(b_1 - t_{1-\frac{\alpha}{2},n-2} \cdot s_{b_1} < \beta_1 < b_1 + t_{1-\frac{\alpha}{2},n-2} \cdot s_{b_1}) = 1-\alpha$$

$$P\left(\beta_1 \in \left(b_1 \mp t_{1-\frac{\alpha}{2},n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}\right)\right) = 1-\alpha$$

Hypotéza, že mezi Y a X není žádný vztah, může být matematicky vyjádřena jako:

H_0: $\beta_1 = 0$

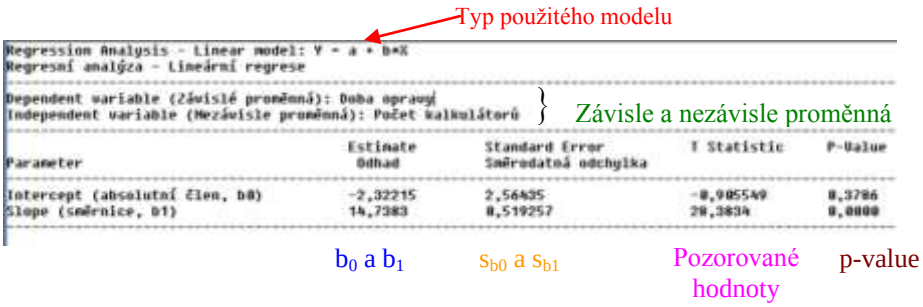
Tato nulová hypotéza se testuje vůči alternativě:

H_A: $\beta_1 \neq 0$

pomocí výše uvedené testové statistiky $\left(\frac{b_1 - \beta_1}{s_{b_1}} \to t_{n-2}\right)$.

14.4.5 Interval spolehlivosti a testy pro β_0

Při konstrukci intervalových odhadů a testování významnosti parametru β_0 postupujeme obdobně jako v případě parametru β_1.



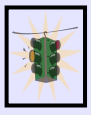
Na základě předpokladu normality popisovaného regresního modelu lze usoudit, že

$$b_0 \to N\left(\beta_0; \sigma_{b_0}^2\right) \;\Rightarrow\; \frac{b_0 - \beta_0}{\sigma_{b_0}} \to N(0;1)$$

A na základě statistického chování rozptylu víme, že

$$\frac{b_0 - \beta_0}{s_{b_0}} \to t_{n-2}$$

Pomocí této výběrové statistiky pak můžeme zkonstruovat **interval spolehlivosti pro β_0**:



Řešený příklad:

$$P\left(\beta_0 \in \left(b_0 \mp t_{1-\frac{\alpha}{2},n-2} \cdot s_{b_0}\right)\right) = 1-\alpha$$

$$P\left(\beta_0 \in \left(b_0 \mp t_{1-\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\left(\frac{1}{n} - \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)}\right)\right) = 1-\alpha$$

Také testování hypotézy o významnosti parametru β_0 se provádí obdobně jako v případě parametru β_1. Souhrnný název pro testy významnosti regresních koeficientu nazýváme **dílčí t-testy**.

Opět se vrátíme k našemu příkladu, vynecháme „ruční výpočet“ a podíváme se, jak pro problematiku dílčích t-testů vypadá výstup statistického software (Statgraphicsu).

Dále v příslušném textovém výstupu nalezneme rovnici vyrovnávací přímky:

```
Doba opravy = -2,32215 + 14,7383*Počet kalkulátorů
```

Z výsledku je patrné, že hypotézu H_0: $\beta_0=0$ nezamítneme s ohledem na hodnotu p-value (0,3786). Na základě toho můžeme prohlásit, že regresní přímka prochází počátkem (absolutní člen regresní přímky můžeme vypustit (považovat za nulový)), což je i logický závěr s ohledem na povahu dat. Druhý z dílčích t-testů nám říká, že směrnice přímky (Slope)

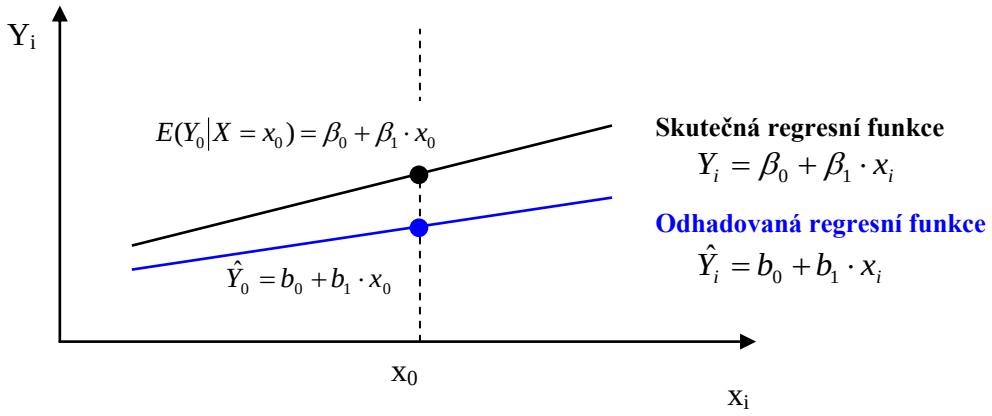
Výklad:

je hodnota, která se významně liší od nuly, neboť jsme zamítli hypotézu H_0: $\beta_1=0$ (p-value=0,0000). Odhadovanou regresní přímku tedy můžeme zapisovat ve tvaru:

Doba opravy = 14,74 . Počet kalkulátorů

14.5 Interval spolehlivosti pro očekávanou hodnotu $E(Y_0 | X=x_0)$

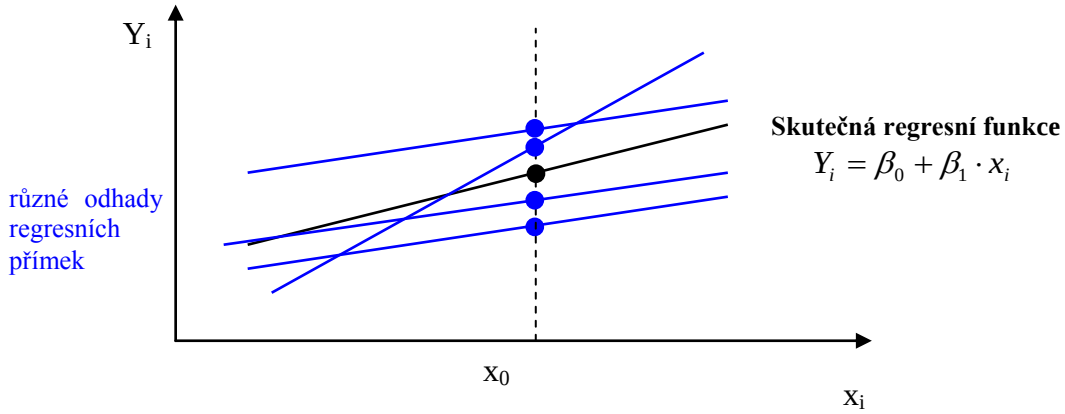
Až dosud jsme studovali aspekty týkající se pozice celé přímky. Nyní se zaměříme na předvídání Y za dané úrovně x.



Jaká je pro daný počet kalkulátoru celková doba opravy? Nejlepším bodovým odhadem této doby je zřejmě bod na odhadované regresní (vyrovnávací) přímce:

$$\hat{Y}_0 = b_0 + b_1 \cdot x_0$$

Víme, že přesnější informaci o odhadované hodnotě $\hat{Y}_0$ nám dá odhad intervalový. Zopakujeme-li výběr, získáme jinou vyrovnávací přímku a tím i jinou hodnotu $\hat{Y}_0$. Všechny hodnoty $\hat{Y}_0$ budou kolísat kolem $E(Y_0|X=x_0)$ a budou znázorňovat rozdělení Y_0.



Bodovým odhadem očekávané hodnoty Y_0 $\left(= E\left(Y_0 \middle| X = x_0\right) = \beta_0 + \beta_1 \cdot x_0\right)$ pro zadanou hodnotu x_0 je statistika:

$$\hat{Y}(x_0) = b_0^* + b_1 \cdot \left(x_0 - \bar{x}\right) = \bar{Y} + \left(\frac{\sum_{i=1}^{n} \left((x_i - \bar{x}) \cdot Y_i\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \cdot (x_0 - \bar{x}) = \sum_{i=1}^{n} \left(\left(\frac{1}{n} + \frac{(x_i - \bar{x}) \cdot (x_0 - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \cdot Y_i \right)$$

Při hledání intervalového odhadu pro $E(Y_0 | X=x_0)$ budeme vycházet zejména z výše odvozené t-statistiky:

$$\frac{\hat{Y}(x_0) - \beta_0 - \beta_1 x_0}{S_{\hat{Y}}} \rightarrow t_{n-2}$$

Z ní na základě běžného postupu, aplikovaného při hledání intervalového odhadu, můžeme získat snadno následující **intervalový odhad pro $E(Y_0 | X=x_0)$**, se spolehlivostí $(1-\alpha)$:

$$P\left(E\left(Y_0 \middle| X = x_0\right) \in \left(\hat{Y}(x_0) - S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}, n-2} ; \hat{Y}(x_0) + S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}, n-2} \right) \right) = 1 - \alpha ,$$

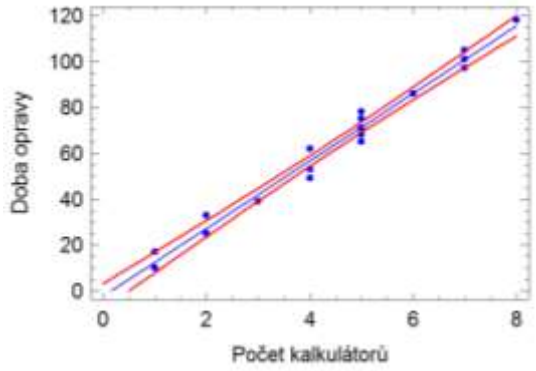
kde

$$\sigma_{\hat{Y}} = \sigma \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{(n-1) \cdot s_x^2} \right)} = \sigma \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \right)} \quad \Rightarrow$$

$$s_{\hat{Y}} = s \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{(n-1) \cdot s_x^2} \right)} = s \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \right)}$$

$$P\left(E\left(Y_0 \middle| X = x_0\right) \in \left(\hat{Y}(x_0) \mp s \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \right)} \cdot t_{1-\frac{\alpha}{2},n-2} \right) \right) = 1 - \alpha$$

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 odhad očekávané hodnoty $E(Y_0 | X=x_0)$ nejpřesnější? Tuto otázku lze zodpovědět nalezením takového x_{opt}, které minimalizuje $S_{\hat{Y}}$:

$$s_{\hat{Y}} = s \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \right)} \qquad \Rightarrow \qquad x_{OPT} = \bar{x}$$

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 $(x_1,..., x_n)$.

14.6 Interval predikce pro jediné pozorování Y_0

V praxi má pro nás mnohdy větší význam tzv. **interval predikce**. Tento interval nám dává odpověď na otázku jaký je interval spolehlivosti Y_0, máme-li k dispozici pouze jediné pozorování na úrovni x_0.

Při predikci Y_0 pak musíme vzít v úvahu:

- Rozptyl odrážející kolísání jednotlivých pozorování, tj. $D\hat{Y}_0$
- Rozptyl odrážející chyby při odhadu vyrovnávací přímky, tj. reziduální rozptyl σ^2

$$DY_0 = D\hat{Y}_0 + \sigma^2 = \sigma^2 \cdot \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \right) + \sigma^2 = \sigma^2 \cdot \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} + 1 \right)$$

Pro lineární regresní model platí, že jednotlivé hodnoty Y_i jsou normálně rozptýleny kolem regresní přímky (e_i mají normální rozdělení), proto:

$$Y_0 \to N\left(\hat{Y}_0; \sigma^2 \cdot \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} + 1 \right) \right) \quad \Rightarrow \quad \frac{Y_0 - \hat{Y}_0}{\sigma \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} + 1 \right)}} \to N(0;1) \quad \Rightarrow$$

$$\frac{Y_0 - \hat{Y}_0}{s \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} + 1 \right)}} \to t_{n-2}$$

Známým způsobem nyní můžeme odvodit **interval predikce**:

$$P\left(Y_0 \in \left(\hat{Y}_0 \mp s \cdot \sqrt{\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} + 1 \right)} \cdot t_{1-\frac{\alpha}{2}, n-2} \right) \right) = 1 - \alpha$$

Pro dostatečně velká n $(n \to \infty)$ se první dva členy pod odmocninou limitně blíží nule a pak je interval predikce:

$$P\left(Y_0 \in \left(\hat{Y}_0 \mp s \cdot t_{1-\frac{\alpha}{2}, n-2} \right) \right) = 1 - \alpha$$

14.7 Index determinace

Pro účely verifikace správnosti zvoleného regresního modelu slouží **index determinace**. 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,

$SS_{\hat{Y}} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$ 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

$$\overline{Y} = \overline{\hat{Y}}$$

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 vysvětlující proměnné), 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

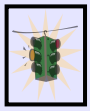
$$R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

a nazveme jej indexem determinace. **Index determinace $\mathbf{R^2}$** tedy udává kvalitu regresního modelu, přesněji řečeno udává, kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno. Tento index 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.

Vyjde-li nízká hodnota indexu determinace , nemusí to ještě znamenat nízký stupeň závislosti mezi proměnnými, ale může to signalizovat chybnou volbu typu regresní funkce.

Hodnoty výše uvedených součtu čtverců prezentuje statistický software většinou ve formě **tabulky ANOVA**, která se vztahuje k testování hypotézy, zda zvolená závislost (statistický software většinou nabízí i jiné typy regrese než lineární) mezi veličinami existuje.

Zatímco dílčí t-testy se používají pro zjištění statistické významnosti jednotlivých regresních koeficientů, hodnota statistiky F-test slouží ke zjištění statistické významnosti těchto koeficientů současně. Soudobá literatura o lineární regresi přitom uvádí, že hodnota statistiky F (tedy společná statistická významnost všech koeficientů jako skupiny) je určující pro významnost jednotlivých koeficientů. To znamená, že bychom se měli nejprve zajímat o hodnotu F-testu, a pokud naznačuje významnost regresních koeficientů jako sady, teprve pak kontrolovat významnost jednotlivých koeficientů. Pokud nejsou koeficienty významné jako sada, je zbytečné zjišťovat významnost u jednotlivých hodnot.

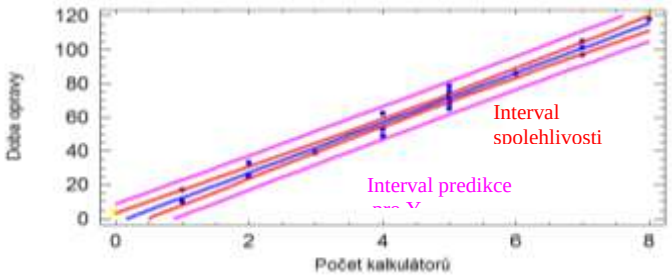


Řešený příklad:

a) Nalezněte 95% pás spolehlivosti a 95% pás predikce kolem regresní přímky pro dobu opravy v závislosti na počtu kalkulátorů (pomocí Statgraphicsu).
b) Nalezněte bodový odhad, intervalový odhad a interval predikce pro očekávanou dobu opravy pěti kalkulátorů.
c) Určete index determinace lineárního regresního modelu pro tento případ
d) Pomocí tabulky ANOVA ověřte, zda skutečně existuje lineární závislost mezi studovanými veličinami.

Řešení:

ada)



adb) Pro $x_0=5$ dostáváme:

Bodový odhad: $$\hat{Y}(5) = b_0 + b_1 \cdot 5 = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x}) \cdot (5 - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \cdot Y_i = 71{,}37$$

Intervalový odhad pro $E(Y_0 | X=x_0)$:

$$P\left(E(Y_0 | X = 5) \in \left(\hat{Y}(5) \mp s \cdot \sqrt{\left(\frac{1}{n} + \frac{(5 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)} \cdot t_{0{,}975, n-2} \right) \right) = 0{,}95 \text{, kde } s = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{n-2}}$$

$$P\left(E(Y_0 | X = 5) \in \langle 69{,}06; 73{,}68 \rangle\right) = 0{,}95$$

Interval predikce:

$$P\left(Y_0 \in \left(\hat{Y}(5) \mp s \cdot \sqrt{\left(\frac{1}{n} + \frac{(5-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + 1\right)} \cdot t_{0,975,n-2}\right)\right) = 0{,}95 \text{, kde } s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y})^2}{n-2}}$$

$$P(Y_0 \in \langle 61{,}59;81{,}15,\rangle) = 0{,}95$$

Predicted Values

X	Predicted Y	95,0% Prediction Limits Lower	Upper	95,0% Confidence Limits Lower	Upper
1,0	12,4161	1,92177	22,9104	7,95981	16,8724
8,0	115,584	105,09	126,078	111,128	120,04
5,0	71,3691	61,5921	81,1462	69,863	73,8752

adc) **Index determinace:**

$$R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{16186{,}44}{16504{,}00} = 0{,}981$$

Index determinace je 98,1%, tzn. že 98,1% celkové doby opravy je vysvětleno lineárním regresním modelem.

add)

Součet čtverců modelu, reziduální a celkový

Výběrový reziduální rozptyl s^2

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	16182,6	1	16182,6	805,62	0,0000
Residual	321,396	16	20,0872		
Total (Corr.)	16504,0	17			

Correlation Coefficient = 0,990215
R-squared = 98,0526 percent
R-squared (adjusted for d.f.) = 97,9309 percent
Standard Error of Est. = 4,48188
Mean absolute error = 3,6428
Durbin-Watson statistic = 1,88411 (P=0,4229)
Lag 1 residual autocorrelation = 0,0172811

Koeficient determinace

Výběrová reziduální směrodatná odchylka

H_0: Mezi celkovou dobou opravy a počtem kalkulátoru neexistuje lineární závislost.
H_A: Mezi celkovou dobou opravy a počtem kalkulátoru existuje lineární závislost.

p-value = 0 $\Rightarrow$ Zamítáme H_0, tzn. lineární závislost považujeme za prokázanou.

Výklad:

14.8 Rozšíření modelu

Odhad regresní funkce, interval spolehlivosti pro $E(Y|X=x_0)$ a interval predikce nám umožňují předvídat Y_0 při **libovolné** hodnotě x_0.

Jestliže $x_0 \in \langle x_1; x_n \rangle$ (x_0 leží mezi pozorovanými hodnotami x_i), proces předvídání se nazývá **interpolace**. Jestliže $x_0 \notin \langle x_1; x_n \rangle$ (x_0 neleží mezi pozorovanými hodnotami x_i), proces předvídání se nazývá **extrapolace**. Vzhledem k tomu, že jak interval spolehlivosti pro $E(Y|X=x_0)$, tak i interval predikce se rozšiřují s rostoucí vzdálenosti od $\bar{x}$, tak čím dále extrapolujeme od pozorovaných hodnot x_i, tím větší riziko podstupujeme. Riziko roste také proto, že mimo interval pozorovaných hodnot nemáme informace o použitelnosti modelu. V podstatě platí, že regresní křivka proložená naměřenými body popisuje chování procesu pouze v rozsahu období, které je těmito body pokryto. Prodloužení regresní křivky mimo toto období (extrapolace) je možné, ale jen do jisté míry a jen s jistým stupněm důvěryhodnosti. My jsme se seznámili s metodami, které umožňují onu důvěryhodnost určit.

Příklad demagogie v regresi:
V civilizovaných zemích klesá dětská úmrtnost a v jistém období lze tento pokles graficky znázornit klesající přímkou. Je zřejmé, že takováto přímka nemůže být libovolně prodloužena. Procento úmrtí prostě nemůže být záporné. V jistém okamžiku se tedy příslušná přímka zalomí v oblouk a časem se zhruba ustálí na nějaké téměř konstantní úrovni.V Británii nastal onen okamžik zlomu v době, kdy začalo hromadné očkování dětí. Pro odpůrce očkování a příslušníky různých extrémních sekt to byl dokonalý statistický důkaz škodlivosti očkování.

Shrnutí:

Často chceme prozkoumat vztah mezi dvěma veličinami, kde jedna z nich, tzv. **nezávisle proměnná x**, má ovlivňovat druhou, tzv. **závisle proměnnou Y**. Předpokládá se, že obě veličiny jsou spojité. Prvním krokem ve zkoumání by mělo být zakreslení dat do bodového grafu, tzv. **korelačního pole** a ověření toho, zda mezi veličinami skutečně existuje předpokládaná závislost, tzv. **regrese.**

Výsledky této části regresní analýzy jsou často na výstupu z počítače prezentovány ve formě **tabulky analýzy rozptylu**.

Nejjednodušší formou regrese je **jednoduchá lineární regrese**, která předpokládá lineární závislost mezi dvěmi veličinami.

Rovnici regresní přímky zapisujeme ve tvaru: $Y_i = \beta_0 + \beta_1 \cdot x_i + e_i$

Odhad regresní přímky nazýváme **vyrovnávací přímka** a zapisujeme jej v jednom z těchto tvarů:

$$\hat{Y}_i = b_0 + b_1 \cdot x_i$$

$$\hat{Y}_i = b_0^* + b_1 \cdot \left(x_i - \bar{x}\right) \qquad \text{(tzv. } \textbf{odchylková forma} \text{ zápisu)}$$

$$\hat{Y}_i = b_0 + b_1 \cdot x_i + e_i$$

(kde e_i označujeme jako chyby predikce (odhadu), resp. rezidua)

Pokud jsou splněny podmínky lineárního regresního modelu, můžeme koeficienty regresní přímky odhadovat **metodou nejmenších čtverců.**

Podmínky lineárního regresního modelu jsou tyto:

$$Y_i = \beta_0 + \beta_1 \cdot x_i + e_i ,$$

kde

1. $E(e_i) = 0$ pro každé $i=1,2,\ldots,n$
   Střední hodnota náhodné složky je nulová.
2. $D(e_i) = \sigma^2$ pro každé $i=1,2,\ldots,n$
   Rozptyl náhodné složky je konstantní.
3. $Cov(e_i, e_j) = 0$ pro každé $i \neq j,$ kde $i, j = 1,2,\ldots,n$
   Kovariance náhodné složky je nulová.
4. **Normalita:** Náhodné složky e_i mají pro $i = 1,2,\ldots,n$ normální rozdělení.
5. **Regresní parametry β_i mohou nabývat libovolných hodnot.**
6. **Regresní model je lineární v parametrech**.

Podmínky lineárníhu regresního modelu je nutno v rámci regresní analýzy ověřit.

Existenci lineárního vztahu mezi dvěma veličinami zjišťujeme tak, že se formálně ptáme, zda je směrnice β_1 rovna nule. Pokud je odpověď na tuto otázku kladná, znamená to, že směrnice vyrovnávací přímky se liší od nuly pouze náhodně, tzn., že vztah mezi sledovanými veličinami není lineární. (Jde o obdobu testu, který je vyhodnocen v tabulce ANOVA.)

Obdobně můžeme testovat významnost absolutního členu vyrovnávací přímky (b_0). Testům významnosti koeficientů vyrovnávací přímky říkáme **dílčí t-testy**.

Intervalový odhad můžeme při regresi hledat jednak pro střední hodnotu Y při dané úrovni x ($E(Y_0| X=x_0)$), jednad pro jednotlivé pozorování (Y_0). Intervalu spolehlivosti pro jednotlivé pozorování říkáme **interval predikce**. Tyto intervalové odhady pro spojitě se měnící hodnoty x tvoří tzv. **pás spolehlivosti kolem regresní přímky,** resp. **pás predikce kolem regresní přímky**.

Kvalitu regresního modelu udává **index determinace R^2**. Přesněji řečeno udává kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno.

Regresní model nám umožňuje provádět rovněž **extrapolaci**, tj. odhad závisle proměnné pro hodnoty nezávisle proměnné ležící mimo interval naměřených hodnot. Extrapolace je vždy spojena s rizikem, že regresní model mimo interval naměřených hodnot pozbývá platnosti.

Otázky

1. Co je to regresní analýza?
2. Vysvětlete pojmy: vysvětlovaná (resp. vysvětlující) proměnná, regresní přímka, vyrovnávací přímka, rezidua.
3. K čemu slouží metoda nejmenších čtverců? Kdy ji nemůžeme použít?
4. Odvoďte metodou nejmenších čtverců koeficienty vyrovnávací přímky.
5. Jaká je interpretace koeficientu β_1?
6. Jakými ukazateli měříme těsnost vzájemné vzájemné vazby? (viz. Náhodný vektor)
7. Čemu říkáme reziduální rozptyl a čím je způsoben?
8. Proč určujeme intervalové odhady koeficientů regresní funkce, resp. proč testujeme významnost koeficientů vyrovnávací přímky?
9. Vysvětlete rozdíl mezi pásem spolehlivosti a pásem predikce.
10. Co je to koeficient determinace?
11. Co je to extrapolace? Jaká jsou její omezení?

Úlohy k řešení

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. Posuďte vhodnost a kvalitu tohoto modelu.

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:

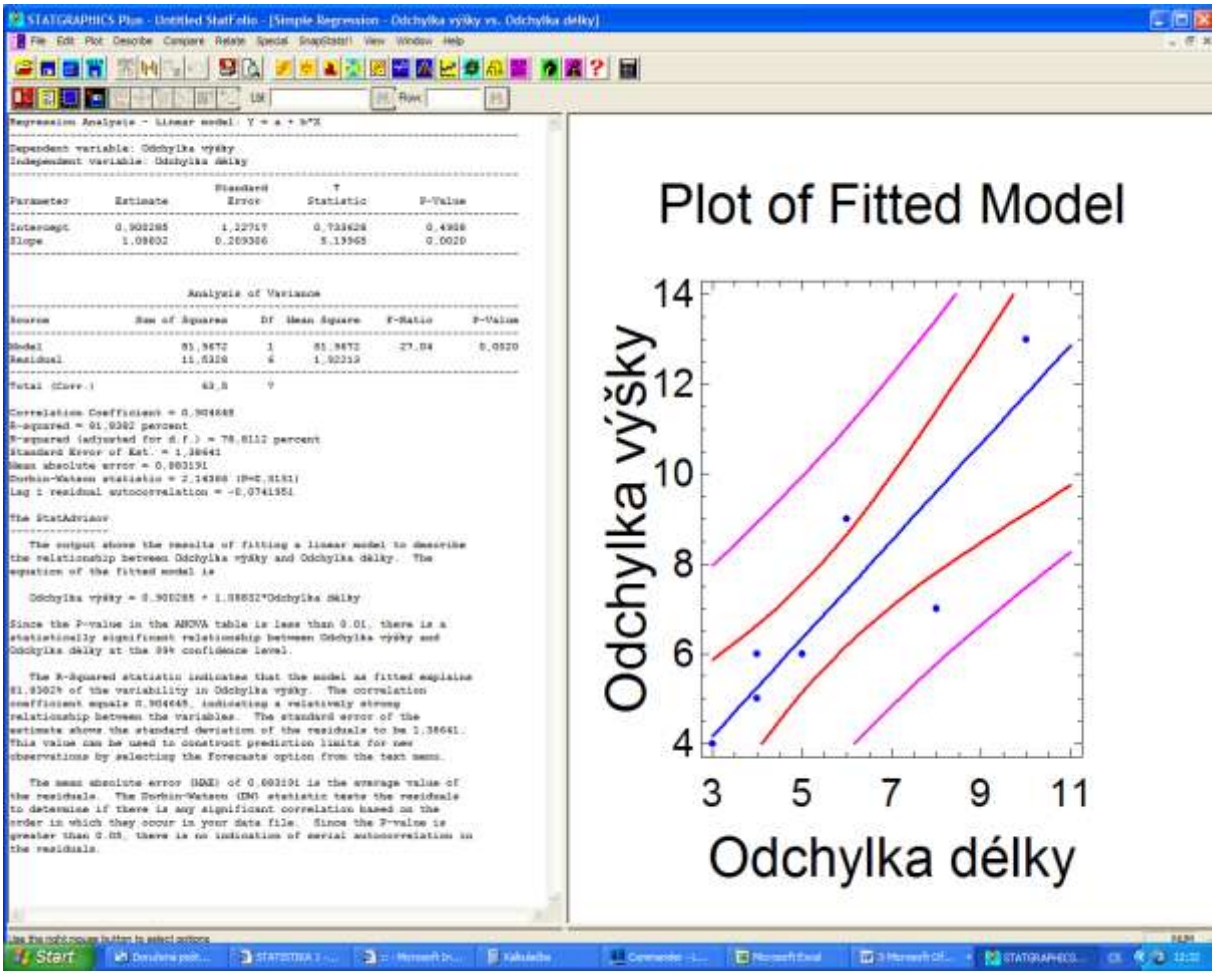
rok	**Šance**	**Morávka**
1931	4,130	2,476
1932	2,386	1,352
1933	2,576	1,238
1934	2,466	1,725
1935	3,576	1,820
1936	2,822	1,913
1937	3,863	2,354
1938	3,706	2,268
1939	3,710	2,534
1940	4,049	2,308
1941	4,466	2,517
1942	2,584	1,726
1943	2,318	1,631
1944	3,721	2,028
1945	3,290	2,423

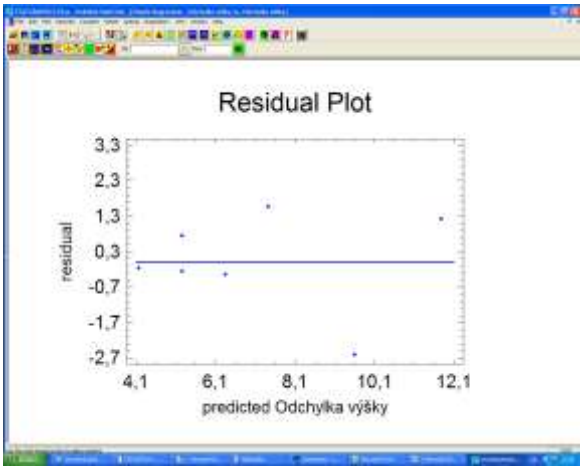
rok	**Šance**	**Moráka**
1946	2,608	1,374
1947	2,045	1,194
1948	3,543	1,799
1949	4,055	2,402
1950	2,224	1,019
1951	2,740	1,552
1952	3,792	1,929
1953	3,087	1,488
1954	1,677	0,803
1955	2,862	1,878
1956	3,802	1,241
1957	2,509	1,165
1958	3,656	1,872
1959	2,447	1,381
1960	2,717	1,679

Za rok 1961 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. (Bodově i intervalově). Zvažte, zda je v tomto případě extrapolace možná.

Řešení:

ad1)





EDF Statistic | Value | Modified Form | P-Value

Kolmogorov-Smirnov D | 0,269882 | 0,84204 | <0.10*
Anderson-Darling A^2 | 0,482561 | 0,544746 | 0.1618*

*Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the currently selected distribution. Other P-values are based on general tables and may be very conservative.

Hypothesis Tests for RESIDUALS

Sample mean = -7,5E-7
Sample median = -0,165242

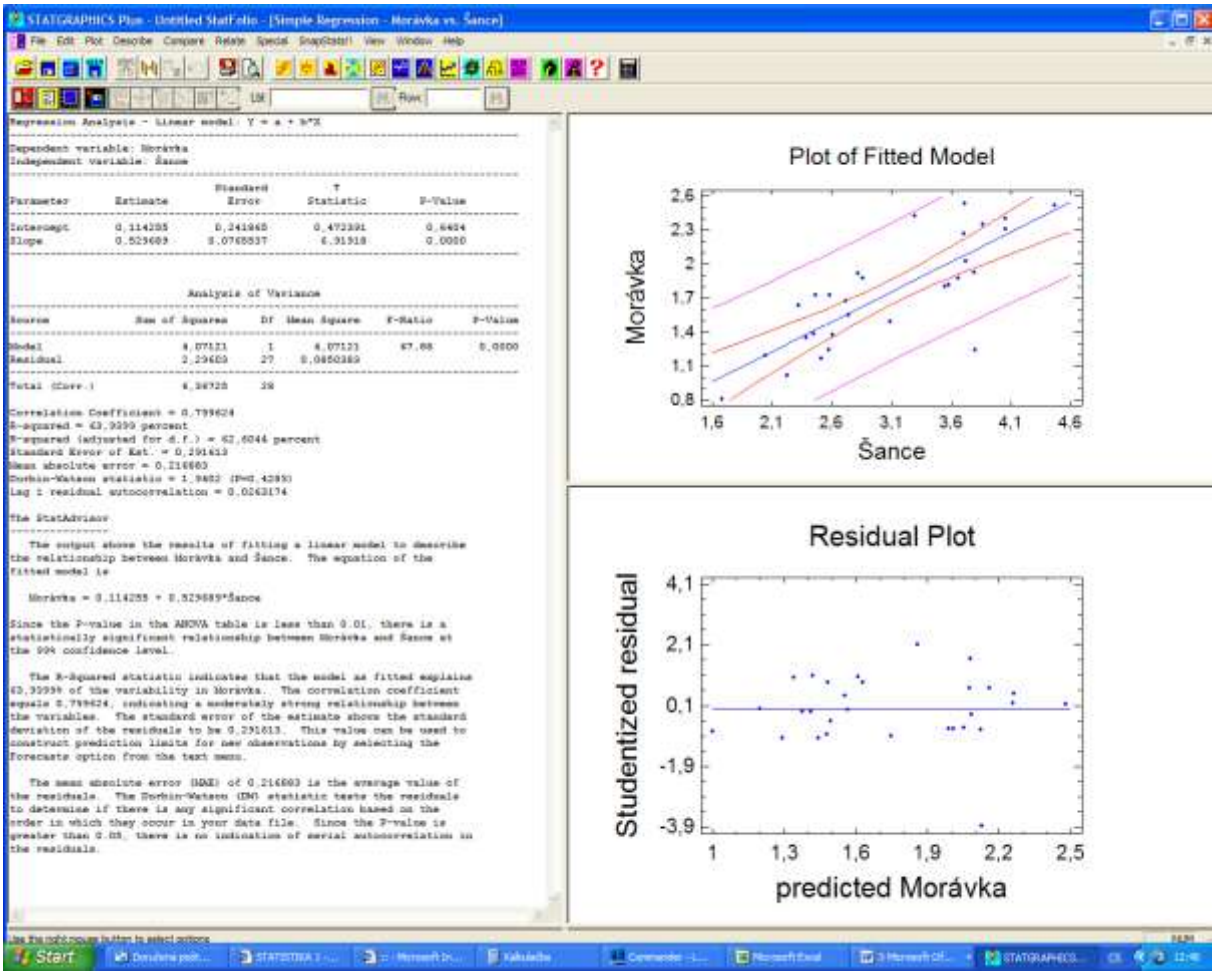
t-test

Null hypothesis: mean = 0,0
Alternative: not equal

Computed t statistic = -0,00000165268
P-Value = 0,999999

Do not reject the null hypothesis for alpha = 0,05.

ad2)



X	Predicted Y	95,00% Prediction Limits Lower	Upper	95,00% Confidence Limits Lower	Upper
1,677	1,00254	0,355341	1,64974	0,755852	1,24923
4,466	2,47984	1,83346	3,12623	2,23531	2,72438
2,91	1,65565	1,0468	2,2648	1,5414	1,76989