Část IV. – Regresní a korelační analýza

Regresní a korelační analýza

- Je známo, že např. hmotnost *m* homogenního tělesa je dána jeho objemem *V*. V tomto případě hovoříme o **funkční závislosti**, tedy $m = f(V)$.
- V mnoha případech je ale třeba zkoumat závislosti, kdy mezi sledovanými znaky (náhodnými proměnnými) neexistuje jednoznačný vztah. V tomto případě hovoříme o **statistické (stochastické) závislosti**.

Regresní a korelační analýza

- K posuzování statistických závislostí slouží regresní a korelační analýza . Úkolem regresní a korelační analýzy je:
  - **Stanovení závislosti** mezi sledovanými kvantitativními znaky (lineární, logaritmická, exponenciální,…), závislost je vyjádřena funkčním předpisem – **regresní analýza**.
  - **Stanovení síly závislosti** mezi sledovanými kvantitativními znaky – **korelační analýza**.

Regresní a korelační analýza

- Sílu **lineární** závislosti mezi dvěma proměnnými můžeme kvantifikovat pomocí Pearsonova (výběrového) korelačního koeficientu:

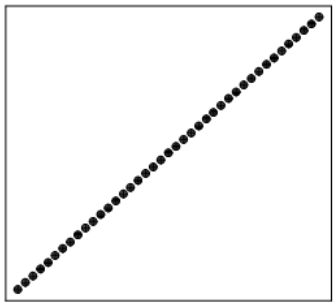
$$r_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\cdot\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \cdot \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} .$$

- Pearsonův korelační koeficient nabývá hodnot z intervalu $\langle -1;1 \rangle$.

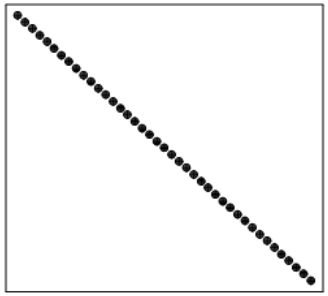
Regresní a korelační analýza

- Předpokladem je, že obě náhodné proměnné, pro které počítáme Pearsonův korelační koeficient, pocházejí z normálního rozdělení.
- Pearsonův korelační koeficient vychází ze vztahu pro výpočet jednoduchého korelačního koeficientu, kde jsou číselné charakteristiky náhodného vektoru (neznámé rozptyly a neznámá kovariance) nahrazeny jejich odhady.

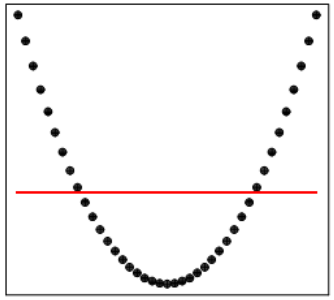
Regresní a korelační analýza



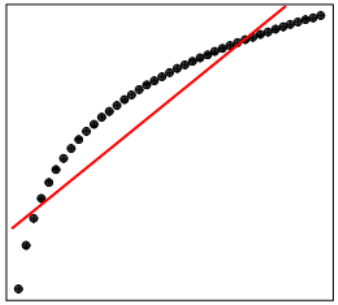
r =1,000



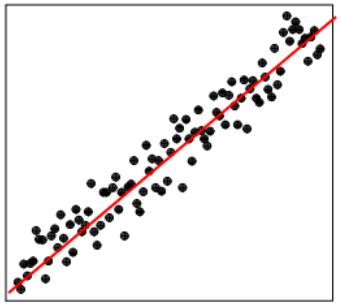
r =-1,000



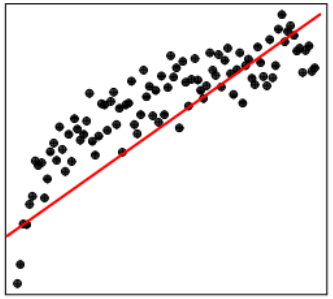
r =0,000



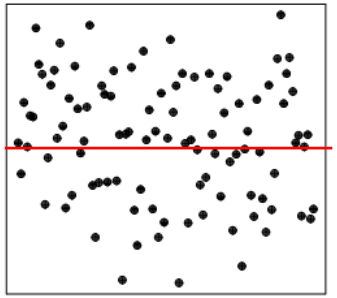
r =0,934



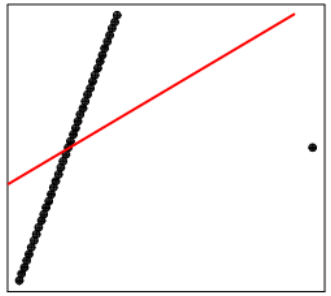
r =0,967



r =0,857



r =-0,143



r =0,608

Regresní a korelační analýza

- V případech, kdy korelační koeficient $r_{X,Y}$ vypočtený z dat získaných náhodným výběrem je blízký nule, má smysl se ptát, zda jsou proměnné X a Y lineárně nezávislé, jinými slovy, zda je hodnota korelačního koeficientu populace $\rho_{X,Y} = 0$.
- Testujeme tedy na základě vypočtené hodnoty Pearsonova korelačního koeficientu, zda je jednoduchý korelační koeficient celé populace rovná nule.

Regresní a korelační analýza

- Nulová hypotéza $H_0: \rho_{X,Y} = 0$ (čili mezi proměnnými X a Y neexistuje lineární vztah).
- V případě alternativní hypotézy má smysl uvažovat tři varianty:
  1. $H_1: \rho_{X,Y} \neq 0$ (oboustranná alternativa, tuto možnost volíme, pokud je vypočtený koeficient korelace blízký 0)

Regresní a korelační analýza

2. $H_1: \rho_{X,Y} > 0$ (pravostranná alternativa, tuto možnost má smysl volit, pokud je vypočtený koeficient korelace větší než 0, výběrový soubor tedy ukazuje na kladnou lineární závislost).
3. $H_1: \rho_{X,Y} < 0$ (levostranná alternativa, tuto možnost má smysl volit, pokud je vypočtený koeficient korelace menší než 0, výběrový soubor tedy ukazuje na zápornou lineární závislost).

Regresní a korelační analýza

- Za předpokladu, že náhodné proměnné X a Y se řídí **normálním** rozdělením pravděpodobnosti, platí pro testovou statistiku:

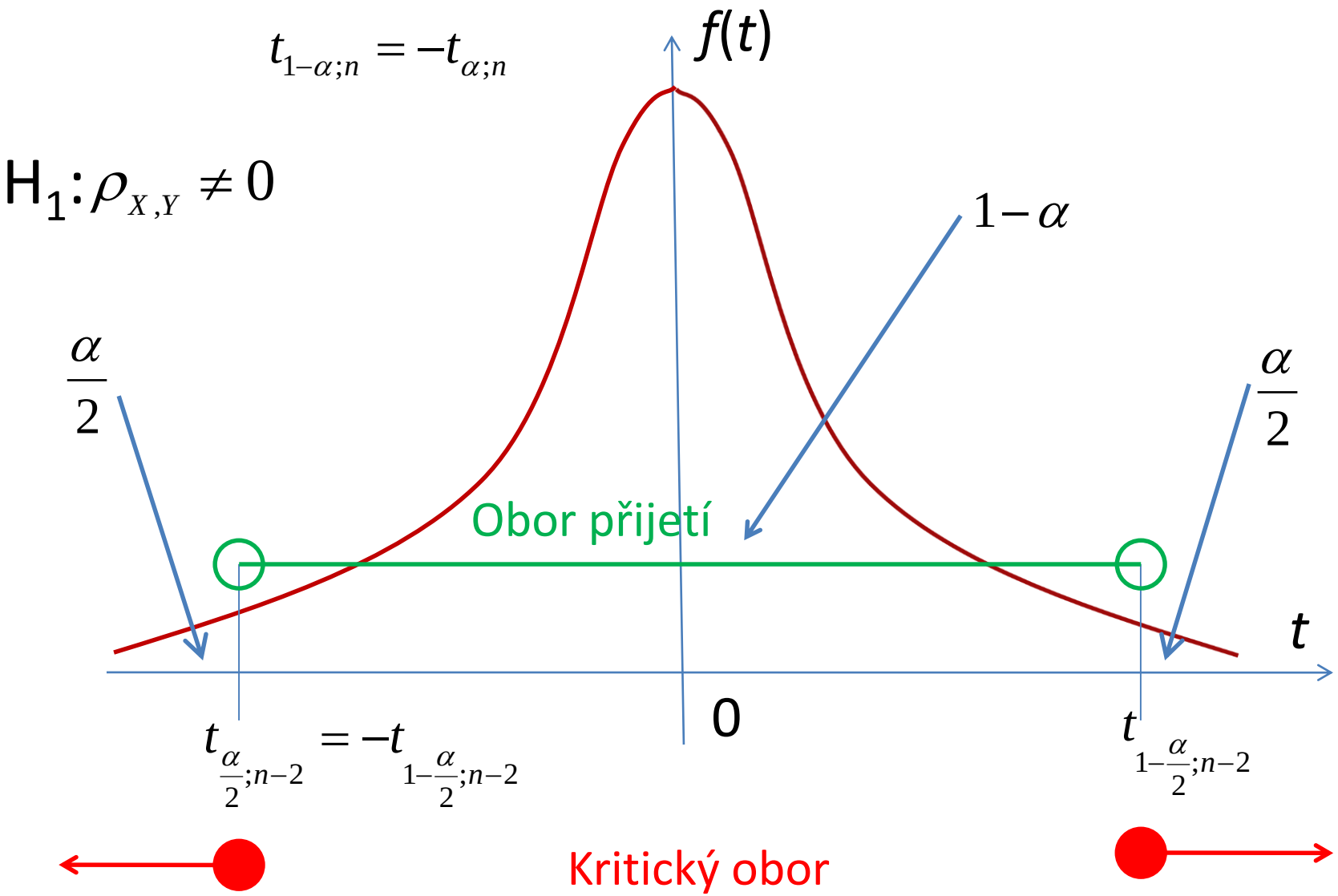
$$T = \frac{r_{X,Y} \cdot \sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} \rightarrow t_{n-2},$$

kde *n* je rozsah výběrového souboru.

Regresní a korelační analýza

- V případě výběru o velkém rozsahu ($n > 30$) lze příslušné Studentovo rozdělení pravděpodobnosti aproximovat normovaným rozdělením pravděpodobnosti N(0,1).
- Při sestrojování kritického oboru a oboru přijetí je nutno vzít v potaz zvolenou alternativní hypotézu.

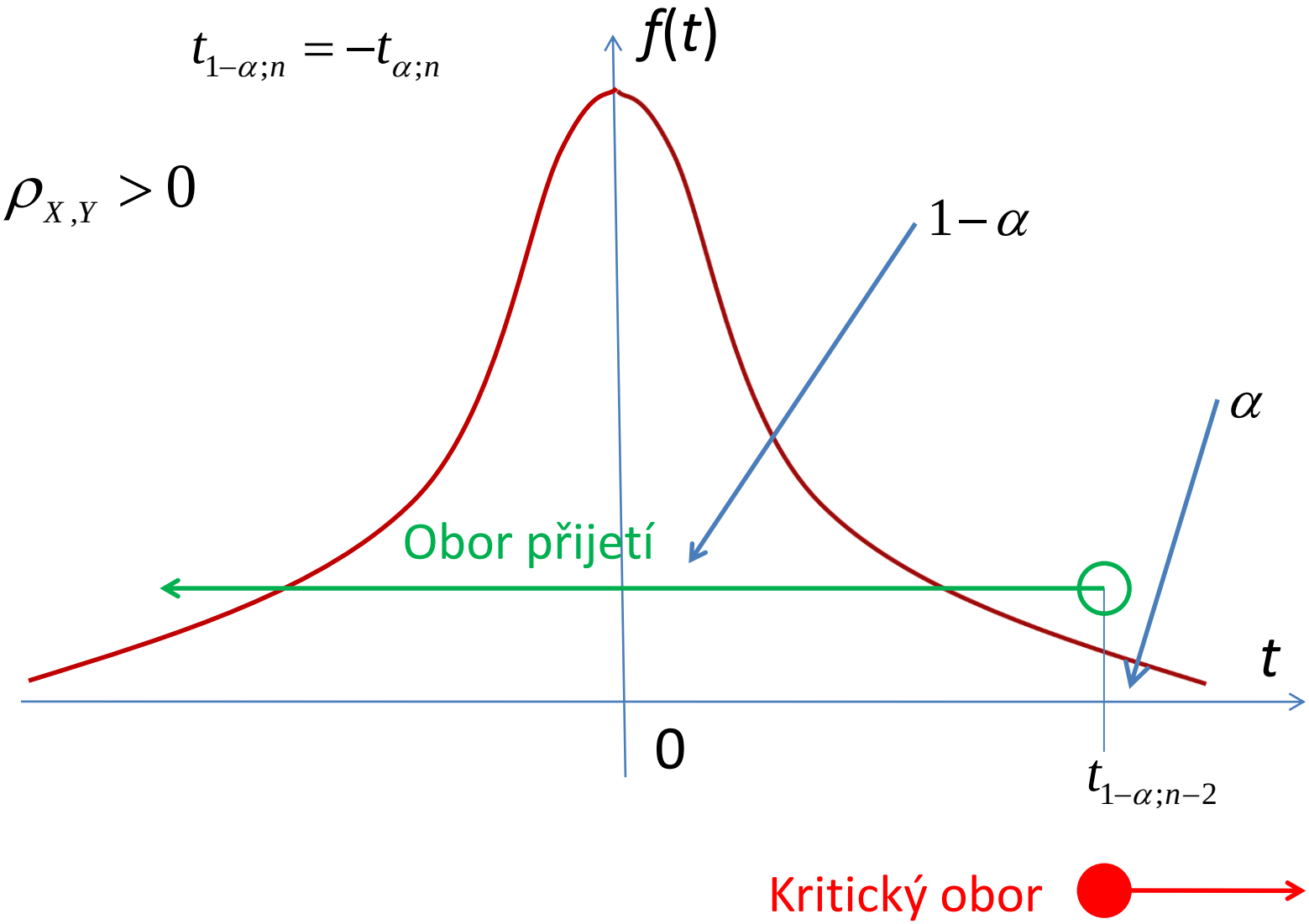
Regresní a korelační analýza



Regresní a korelační analýza

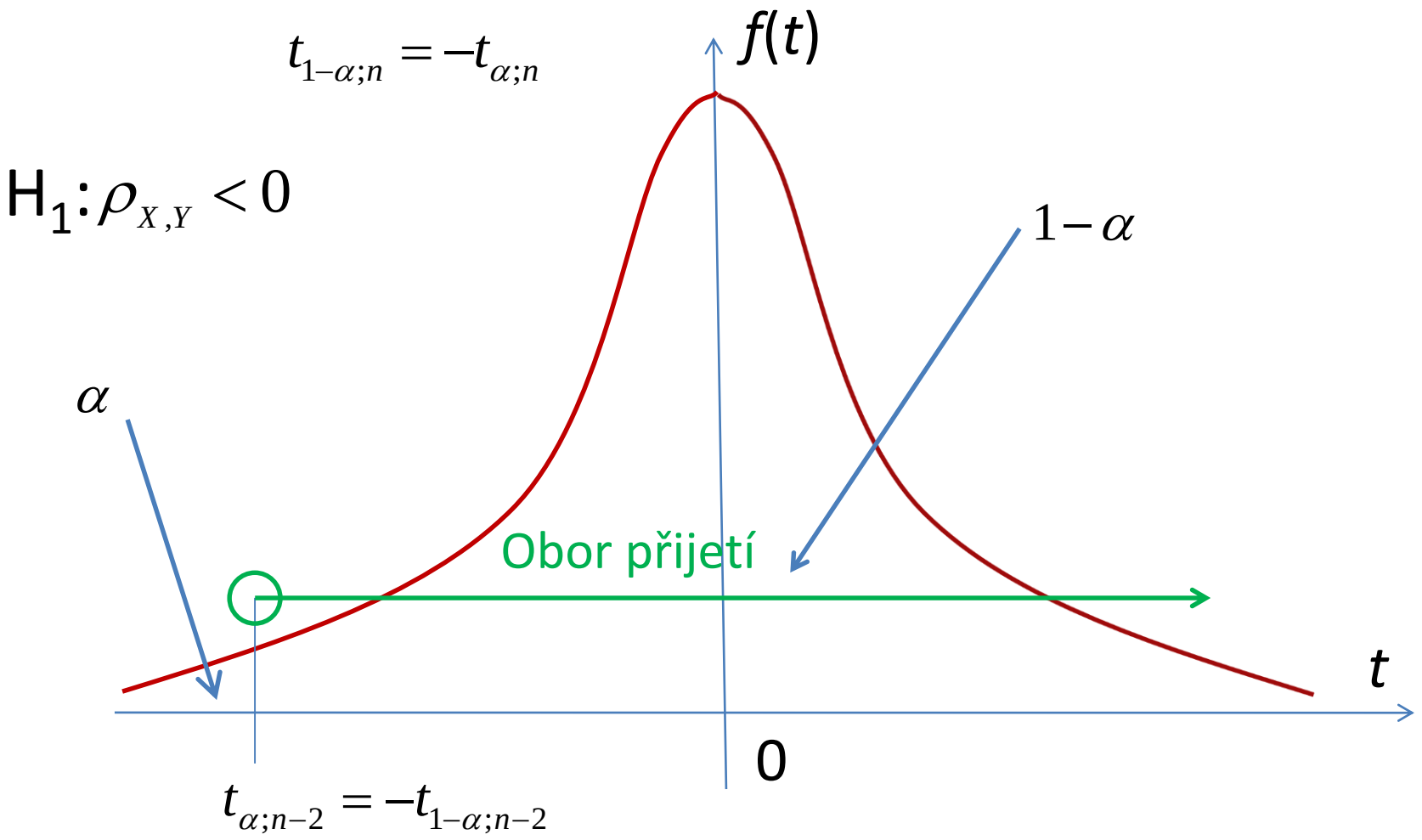
$t_{1-\alpha;n} = -t_{\alpha;n}$

$H_1: \rho_{X,Y} > 0$



Kritický obor

Regresní a korelační analýza



Kritický obor

Regresní a korelační analýza

- Výsledek testu:
  - Leží-li vypočtená hodnota testové statistiky x_{obs} v oboru přijetí, potom nezamítáme nulovou hypotézu o lineární nezávislosti proměnných X a Y.
  - Leží-li vypočtená hodnota testové statistiky x_{obs} v kritickém oboru, potom zamítáme nulovou hypotézu ve prospěch alternativní hypotézy.

Regresní a korelační analýza

- **Př.:** V náhodném výběru o rozsahu 25 pozorování byl vypočítán koeficient korelace $r_{X,Y} = 0{,}23$. Na hladině významnosti 0,05 otestujte, zda lze na základě tohoto výsledku usuzovat na lineární nezávislost mezi proměnnými X a Y v celé populaci.

Regresní a korelační analýza

- Nulová hypotéza $H_0: \rho_{X,Y} = 0$ (čili mezi proměnnými X a Y neexistuje lineární vztah).
- V případě alternativní hypotézy má smysl uvažovat dvě varianty:
  1. $H_1: \rho_{X,Y} \neq 0.$
  2. $H_1: \rho_{X,Y} > 0.$

Regresní a korelační analýza

- Výpočet pozorované hodnoty testové statistiky:

$$x_{obs} = \frac{r_{X,Y} \cdot \sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} = \frac{0{,}23 \cdot \sqrt{25-2}}{\sqrt{1-0{,}23^2}} = 1{,}133.$$

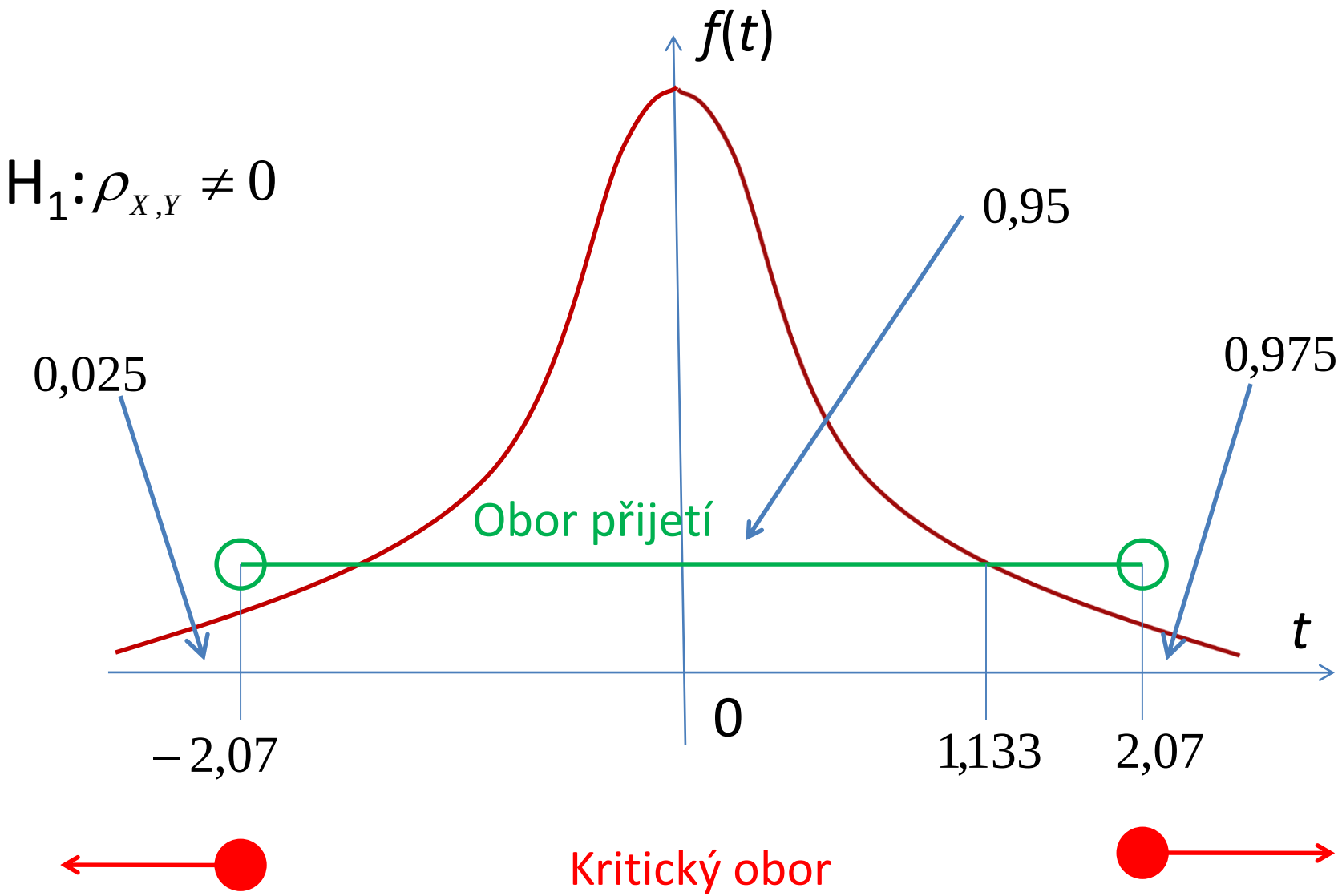
- Příslušné kvantily Studentova rozdělení získáme z tabulek:

$$t_{1-\frac{\alpha}{2};n-2} = t_{1-0{,}025;25-2} = t_{0{,}975;23} \doteq 2{,}07, t_{\frac{\alpha}{2};n-2} = -t_{1-\frac{\alpha}{2};n-2} = -2{,}07,$$

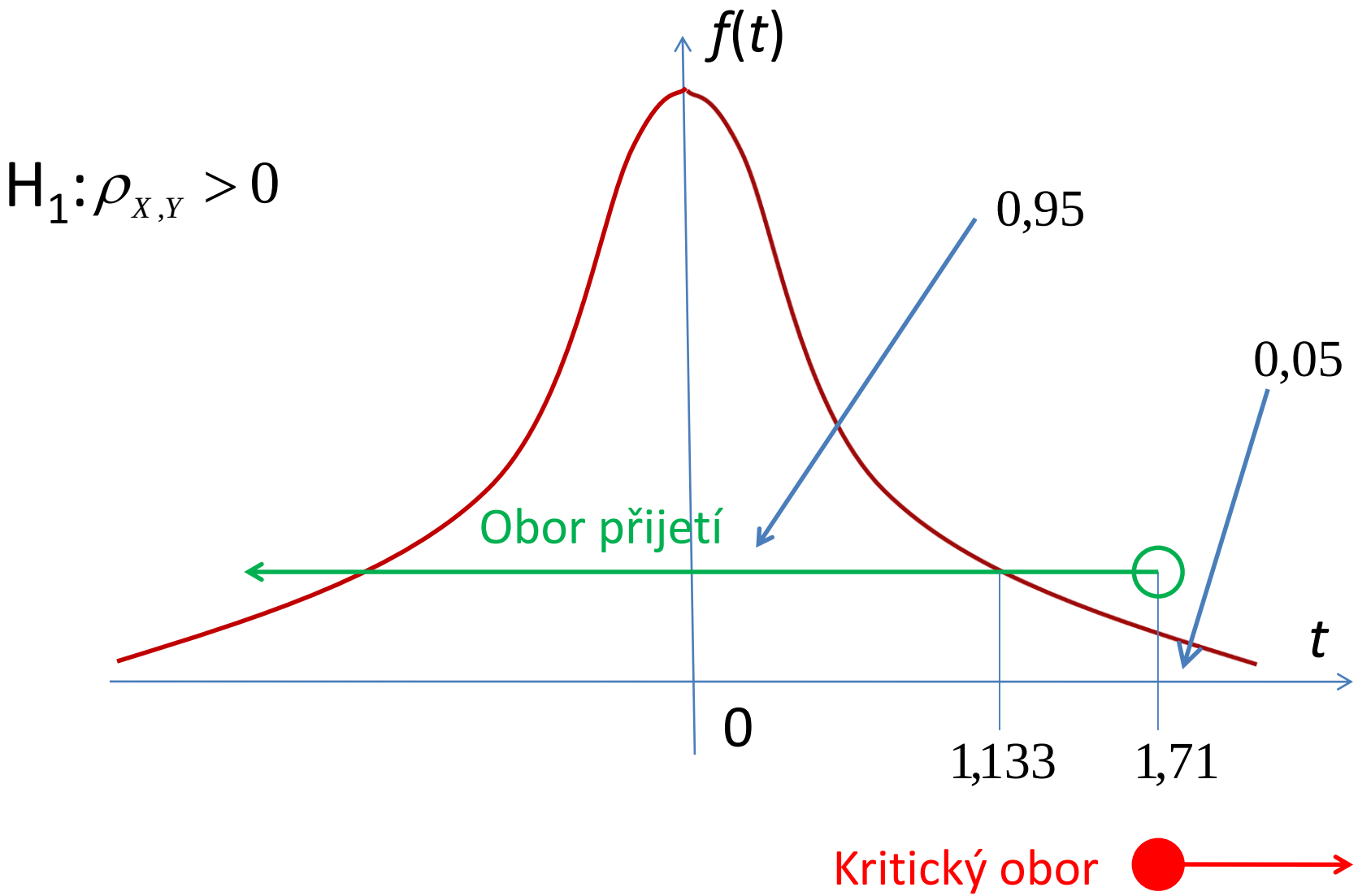
$$t_{1-\alpha;n-2} = t_{1-0{,}05;25-2} = t_{0{,}95;23} \doteq 1{,}71.$$

Stupně volnosti	$t_{0.75}$	$t_{0.9}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$	$t_{0.9975}$	$t_{0.999}$	$t_{0.9995}$
1	1,00	3,08	6,31	12,71	31,82	63,66	127,32	318,31	636,62
2	0,82	1,89	2,92	4,30	6,96	9,92	14,09	22,33	31,60
3	0,76	1,64	2,35	3,18	4,54	5,84	7,45	10,21	12,92
4	0,74	1,53	2,13	2,78	3,75	4,60	5,60	7,17	8,61
5	0,73	1,48	2,02	2,57	3,36	4,03	4,77	5,89	6,87
6	0,72	1,44	1,94	2,45	3,14	3,71	4,32	5,21	5,96
7	0,71	1,41	1,89	2,36	3,00	3,50	4,03	4,79	5,41
8	0,71	1,40	1,86	2,31	2,90	3,36	3,83	4,50	5,04
9	0,70	1,38	1,83	2,26	2,82	3,25	3,69	4,30	4,78
10	0,70	1,37	1,81	2,23	2,76	3,17	3,58	4,14	4,59
11	0,70	1,36	1,80	2,20	2,72	3,11	3,50	4,02	4,44
12	0,70	1,36	1,78	2,18	2,68	3,05	3,43	3,93	4,32
13	0,69	1,35	1,77	2,16	2,65	3,01	3,37	3,85	4,22
14	0,69	1,35	1,76	2,14	2,62	2,98	3,33	3,79	4,14
15	0,69	1,34	1,75	2,13	2,60	2,95	3,29	3,73	4,07
16	0,69	1,34	1,75	2,12	2,58	2,92	3,25	3,69	4,01
17	0,69	1,33	1,74	2,11	2,57	2,90	3,22	3,65	3,97
18	0,69	1,33	1,73	2,10	2,55	2,88	3,20	3,61	3,92
19	0,69	1,33	1,73	2,09	2,54	2,86	3,17	3,58	3,88
20	0,69	1,33	1,72	2,09	2,53	2,85	3,15	3,55	3,85
21	0,69	1,32	1,72	2,08	2,52	2,83	3,14	3,53	3,82
22	0,69	1,32	1,72	2,07	2,51	2,82	3,12	3,50	3,79
23	0,69	1,32	1,71	2,07	2,50	2,81	3,10	3,48	3,77
24	0,68	1,32	1,71	2,06	2,49	2,80	3,09	3,47	3,75
25	0,68	1,32	1,71	2,06	2,49	2,79	3,08	3,45	3,73
26	0,68	1,31	1,71	2,06	2,48	2,78	3,07	3,43	3,71
27	0,68	1,31	1,70	2,05	2,47	2,77	3,06	3,42	3,69
28	0,68	1,31	1,70	2,05	2,47	2,76	3,05	3,41	3,67
29	0,68	1,31	1,70	2,05	2,46	2,76	3,04	3,40	3,66
30	0,68	1,31	1,70	2,04	2,46	2,75	3,03	3,39	3,65
40	0,68	1,30	1,68	2,02	2,42	2,70	2,97	3,31	3,55
50	0,68	1,30	1,68	2,01	2,40	2,68	2,94	3,26	3,50
60	0,68	1,30	1,67	2,00	2,39	2,66	2,91	3,23	3,46
70	0,68	1,29	1,67	1,99	2,38	2,65	2,90	3,21	3,44
80	0,68	1,29	1,66	1,99	2,37	2,64	2,89	3,20	3,42
100	0,68	1,29	1,66	1,98	2,36	2,63	2,87	3,17	3,39
120	0,68	1,29	1,66	1,98	2,36	2,62	2,86	3,16	3,37
∞	0,67	1,28	1,64	1,96	2,33	2,58	2,81	3,09	3,29

Regresní a korelační analýza



Regresní a korelační analýza



Regresní a korelační analýza

- V obou případech vidíme, že pozorovaná hodnota testového kritéria leží v oboru přijetí, výsledkem tedy je konstatování, že nezamítáme nulovou hypotézu, můžeme tedy předpokládat, že náhodné proměnné jsou lineárně nezávislé.

Regresní a korelační analýza

- V případech, kdy není splněna normalita obou náhodných výběrů, lze místo Pearsonova korelačního koeficientu použít Spearmanův korelační koeficient.
- Mějme náhodný výběr z dvourozměrného rozdělení $(X_1,Y_1),\ldots,(X_n,Y_n)$. Zaveďme nyní $P_1,\ldots,P_n$ jako pořadí veličiny $X_1,\ldots,X_n$ a $R_1,\ldots,R_n$ jako pořadí veličiny $Y_1,\ldots,Y_n$.

Regresní a korelační analýza

- V případě, že máme několik stejných hodnot, potom jim přiřadíme průměrné pořadí.
- Je zřejmé, že pokud s rostoucím X_i bude růst i Y_i, potom bude stejný vztah platit i pro jejich pořadí.
- Pokud s klesajícím X_i bude klesat i Y_i, potom bude stejný vztah platit i pro jejich pořadí.
- Budou-li veličiny X a Y nezávislé, potom budou i hodnoty jejich pořadí náhodně přeházené.

Regresní a korelační analýza

- Spearmanův korelační koeficient r_S je potom definován vztahem:

$$r_s = 1 - \frac{6}{n \cdot \left(n^2 - 1\right)} \cdot \sum_{i=1}^{n} \left(P_i - R_i\right)^2 .$$

- Spearmanův korelační koeficient nabývá hodnot z intervalu $\langle -1,1 \rangle$.

Regresní a korelační analýza

- Při shodném pořadí nabývá hodnota Spearmanova korelačního koeficientu hodnoty 1.
- Při opačném pořadí nabývá hodnoty -1.
- V případě nezávislosti obou veličin X a Y nabývá hodnoty 0.

Regresní a korelační analýza

- Pokud se v náhodném výběru vyskytuje mnoho shod (tj. stejně velkých pozorování), potom se doporučuje používat korigovaný Spearmanův koeficient. Zaveďme:
  - Veličinu t_x jako počty stejných hodnot proměnné X.
  - Veličinu t_y jako počty stejných hodnot proměnné Y.

Regresní a korelační analýza

- Potom korigovaný Spearmanův koeficient definujeme vztahem:

$$r_{s_{korig}} = 1 - \frac{6}{n^3 - n - T_X - T_Y} \cdot \sum_{i=1}^{n} \left(P_i - R_i\right)^2 ,$$

kde $T_X = \frac{1}{2} \cdot \sum_x \left(t_x^3 - t_x\right)$ a $T_Y = \frac{1}{2} \cdot \sum_y \left(t_y^3 - t_y\right)$.

Regresní a korelační analýza

- Vyjde-li hodnota Spearmanova korelačního koeficientu blízká nule, může nás zase zajímat odpověď na otázku, zda je jeho hodnota statisticky významná, jinými slovy zda lze veličiny X a Y považovat za nezávislé. Dostáváme následující hypotézy:
  - H_0 – veličiny X a Y jsou nezávislé náhodné veličiny.
  - H_1 – veličiny X a Y jsou závislé náhodné veličiny.

Regresní a korelační analýza

- Testovou statistikou je absolutní hodnota Spearmanova korelačního koeficientu, tedy:

$$x_{obs} = |r_S|.$$

- Nulovou hypotézu zamítáme v tom případě, pokud platí, že:

$$x_{obs} \geq r_s^*(\alpha, n),$$

kde $r_s^*(\alpha, n)$ je pro $n \leq 30$ tabelovaná hodnota.

Regresní a korelační analýza

n	α=0,05	α=0,01	n	α=0,05	α=0,01
5	0,900		18	0,399	0,564
6	0,829	0,943	19	0,388	0,549
7	0,714	0,893	20	0,377	0,534
8	0,643	0,833	21	0,368	0,521
9	0,600	0,783	22	0,359	0,508
10	0,564	0,745	23	0,351	0,496
11	0,523	0,736	24	0,343	0,485
12	0,497	0,703	25	0,336	0,475
13	0,475	0,673	26	0,329	0,465
14	0,457	0,646	27	0,323	0,456
15	0,441	0,623	28	0,317	0,448
16	0,425	0,601	29	0,311	0,440
17	0,412	0,582	30	0,305	0,432

Regresní a korelační analýza

- Pro *n*>30 se kritická hodnota $r_s^*(\alpha, n)$ stanoví:

$$r_s^*(\alpha, n) = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-1}},$$

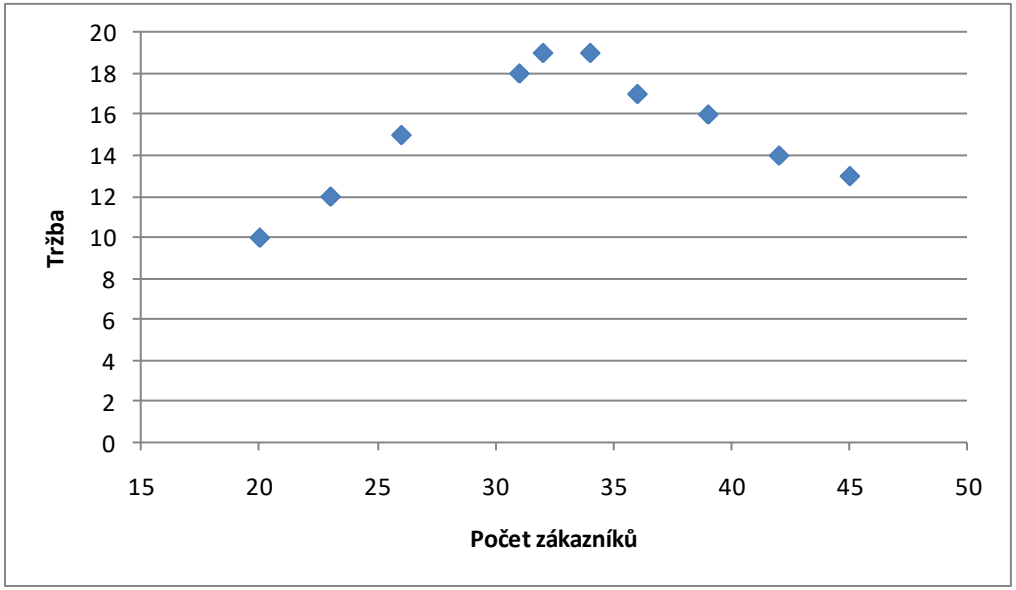
kde v čitateli je příslušný kvantil normovaného normálního rozdělení (jeho hodnotu např. nalezneme ve statistických tabulkách).

Regresní a korelační analýza

- **Př.:** V obchodě, zabývajícím se prodejem náhradních dílů do automobilů, bylo provedeno měření počtu zákazníků přicházejících do obchodu za 1 hodinu a odpovídajících tržeb za 1 hodinu vyjádřených v tisících Kč. Stanovte hodnotu Spearmanova korelačního koeficientu a pro α=0,05 otestujte hypotézu, zda lze počet přicházejících zákazníků za hodinu a hodinové tržby považovat za nezávislé veličiny.

Regresní a korelační analýza

Počet zákazníků - X_i	Hodinová tržba - Y_i
20	10
23	12
26	15
31	18
32	19
34	19
36	17
39	16
42	14
45	13



Regresní a korelační analýza

- Nejdříve musíme jednotlivým hodnotám veličin X a Y přiřadit pořadí.

Počet zákazníků - X_i	Hodinová tržba - Y_i	Pořadí P_i	Pořadí R_i	$(P_i-R_i)^2$
20	10	1	1	0
23	12	2	2	0
26	15	3	5	4
31	18	4	8	16
32	19	5	9,5	20,25
34	19	6	9,5	12,25
36	17	7	7	0
39	16	8	6	4
42	14	9	4	25
45	13	10	3	49
∑				130,5

Regresní a korelační analýza

- Nyní můžeme dosadit do vztahu pro výpočet Spearmanova korelačního koeficientu:

$$r_s = 1 - \frac{6}{n \cdot \left(n^2 - 1\right)} \cdot \sum_{i=1}^{n} \left(P_i - R_i\right)^2 = 1 - \frac{6}{10 \cdot \left(10^2 - 1\right)} \cdot 130{,}5 \doteq 0{,}21.$$

- Nyní budeme testovat hypotézu o nezávislosti obou veličin.

Regresní a korelační analýza

- H_0 – Počet přicházejících zákazníků za hodinu a hodinové tržby obchodu jsou nezávislé veličiny.
- H_1 – Počet přicházejících zákazníků za hodinu a hodinové tržby obchodu jsou závislé veličiny.
- Z tabulky odečteme kritickou hodnotu testu pro n=10 (máme 10 pozorování) a hladinu významnosti α=0,05), která je rovna 0,564.

Regresní a korelační analýza

- Porovnáním pozorovaného hodnoty testové statistiky (absolutní hodnota Spearmanova korelačního koeficientu) s kritickou hodnotou testu vidíme, že nezamítáme nulovou hypotézu o nezávislosti obou veličin.

Regresní a korelační analýza

- **Lineární regrese** – závislost proměnných je vyjádřena funkcí lineární v parametrech (resp. se dá na funkci lineární v parametrech převést vhodnou transformací) – např. $Y = \beta_0 + \beta_1 \cdot x$.
- **Nelineární regrese** – závislost proměnných je vyjádřena funkcí nelineární v parametrech (a ani nelze na funkci lineární v parametrech převést pomocí žádné transformace) – např. $Y = \beta_0 \cdot \beta_1 \cdot x$.

Regresní a korelační analýza

- **Jednoduchá regrese** – studuje závislost jedné proměnné na druhé proměnné.
- **Vícenásobná regrese** – studuje závislost jedné proměnné na několika proměnných.

Regresní a korelační analýza

- **Vysvětlovaná (závisle) proměnná Y** – proměnná, jejíž chování se snažíme vysvětlit, tedy popsat vyrovnávací křivkou.
- **Vysvětlující (nezávisle) proměnná x** – proměnná, jejíž chování vysvětluje chování závisle proměnné Y. Tato proměnná je příčinnou proměnnou, v důsledku její změny se mění vysvětlovaná proměnná.

Regresní a korelační analýza

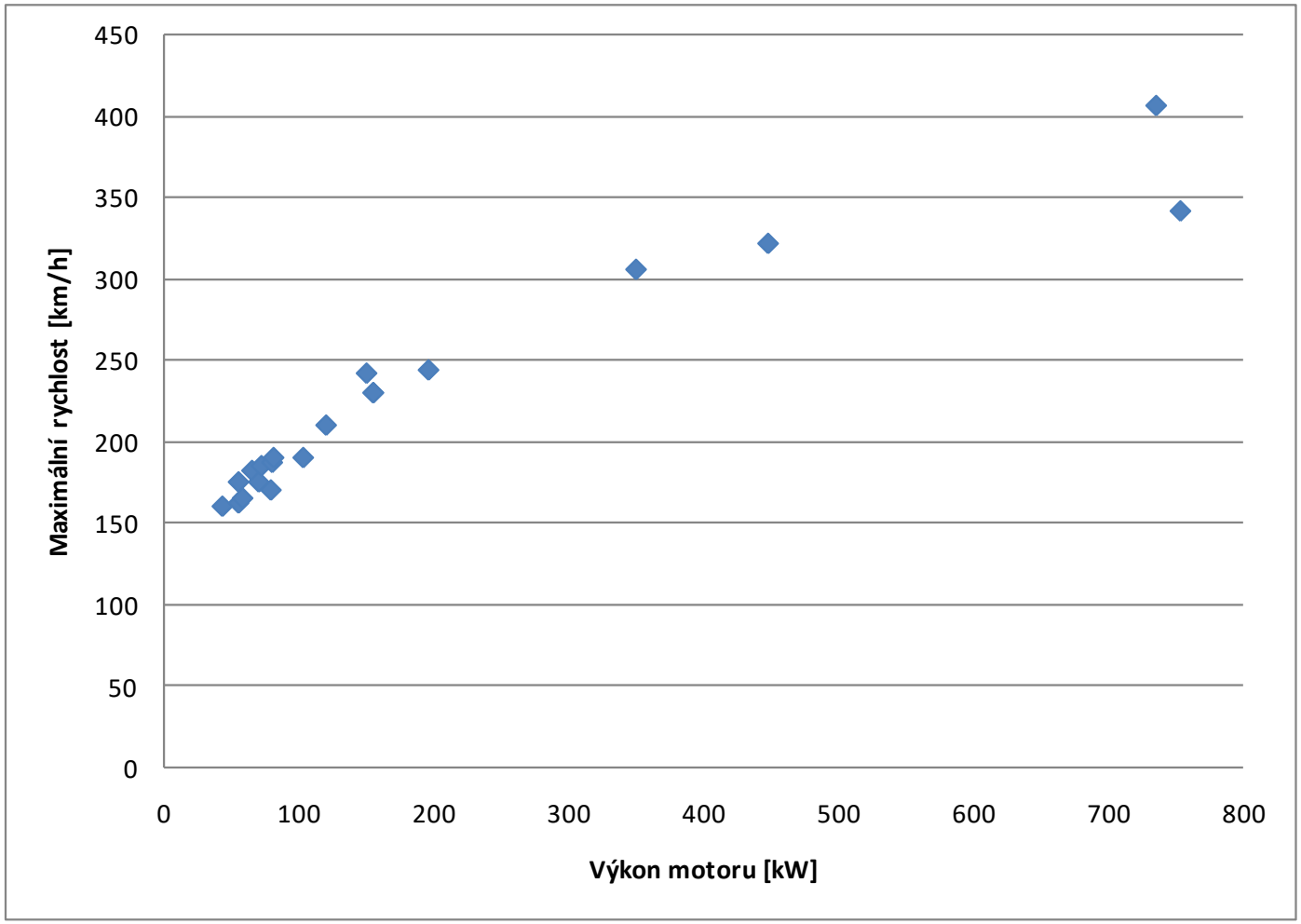
Zajímá nás, zda existuje nějaká závislost mezi výkonem motoru automobilu a jeho maximální rychlostí. Výkon motoru je v tomto případě vysvětlující proměnná a maximální rychlost je vysvětlovaná proměnná.

Výkon motoru [kW]	Maximální rychlost [km/h]
43	160
55	162
55	175
58	165
65	182
70	175
72	185
79	170
80	187
81	190
103	190
120	210
150	242
155	230
155	230
196	244
350	306
448	322
736	407
754	342

Regresní a korelační analýza

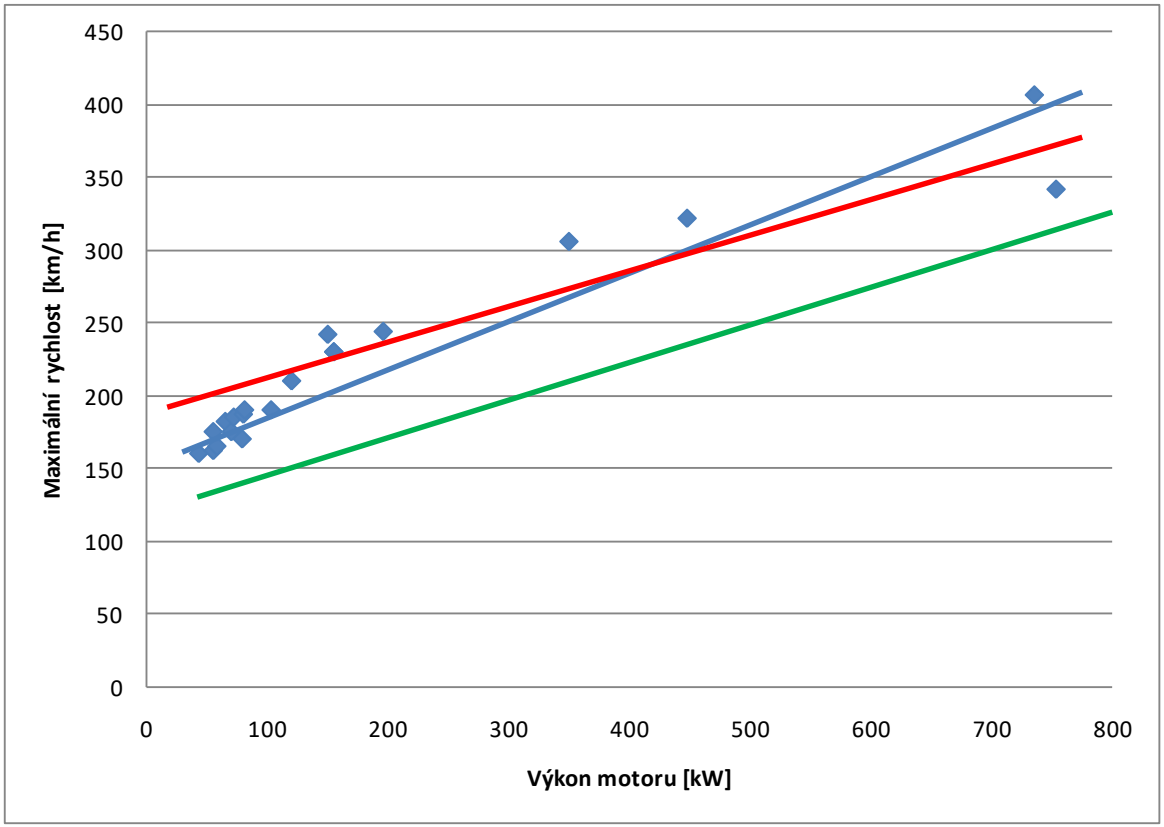
- Orientačně („podle oka“) lze druh a sílu závislosti mezi vysvětlující a vysvětlovanou proměnnou posoudit na základě bodového grafu $[x_i, Y_i]$ – **korelační pole**.
- Dále se budeme podrobně zabývat pouze jednoduchou lineární regresí, vyrovnávací křivka má tvar přímky.

Regresní a korelační analýza



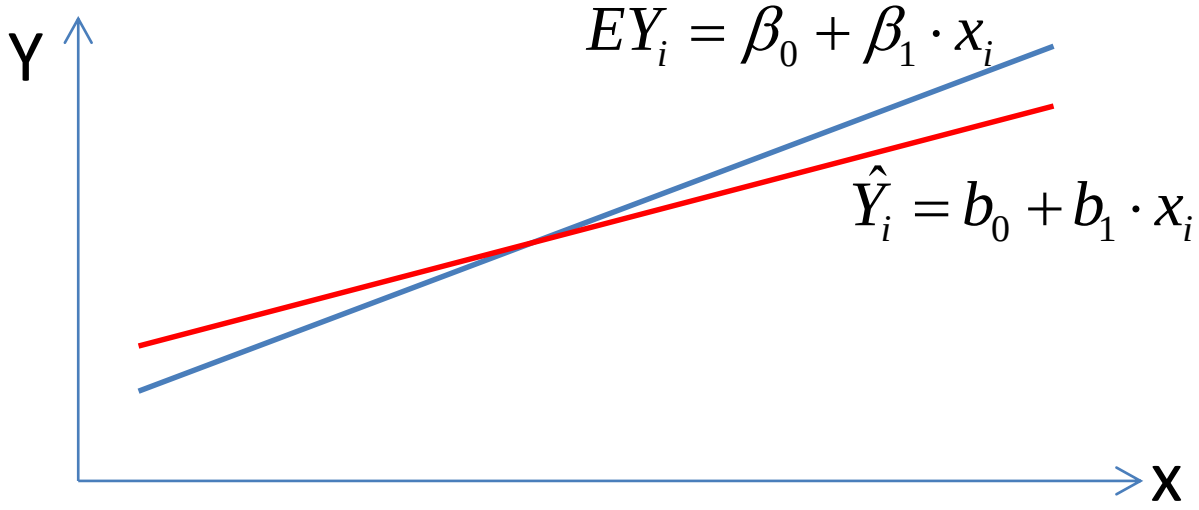
Regresní a korelační analýza

- Otázkou je, jak jednotlivými body proložit vyrovnávací křivku.



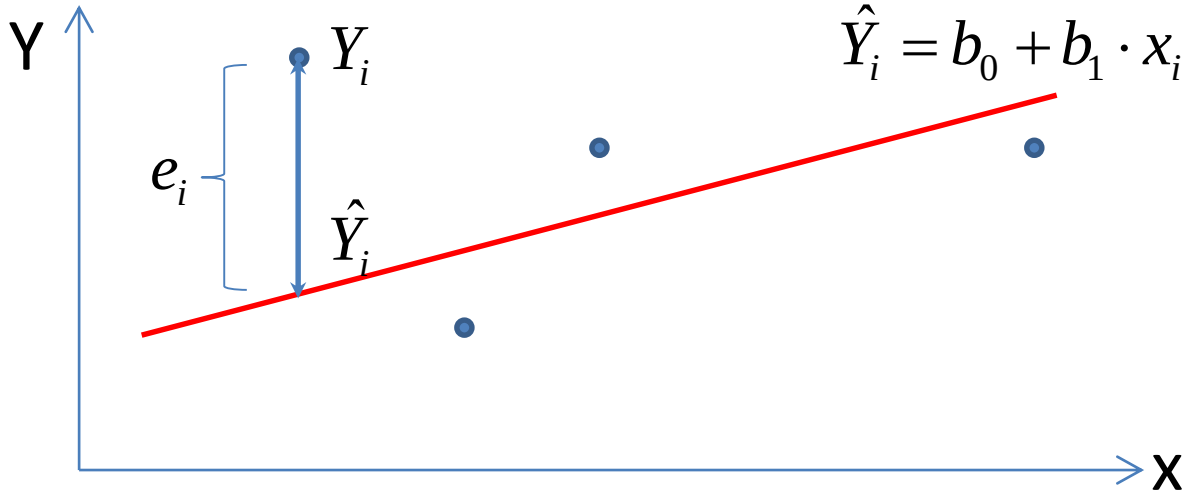
Regresní a korelační analýza

- Regresní funkce – $EY_i = \beta_0 + \beta_1 \cdot x_i$ – skutečná regrese **populace**, v praxi neznámá, proto regresní funkci pouze odhadujeme, zapisujeme tedy $\hat{Y}_i = b_0 + b_1 \cdot x_i$.



Regresní a korelační analýza

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



Regresní a korelační analýza

- Úkolem je najít vyrovnávací křivku $\hat{Y}_i = b_0 + b_1 \cdot x_i$ takovou, abychom získali co nejméně rozptýlený soubor reziduí. Můžeme tedy minimalizovat:
  - Součet reziduí $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)$. ✗
  - Součet absolutních odchylek reziduí $\sum_{i=1}^{n} \left| Y_i - \hat{Y} \right|$. ✗
  - Součet druhých mocnin reziduí $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$. ✓

Regresní a korelační analýza

- K nalezení koeficientů vyrovnávací přímky tedy použijeme **metodu nejmenších čtverců**.
- Pro zjednodušení nejdříve upravme vztah pro $\hat{Y}_i$ do vhodnější formy – tzv. odchylková forma:

$$\hat{Y}_i = b_0 + b_1 \cdot x_1 = \left(b_0 + b_1 \cdot \bar{x}\right) + b_1 \cdot \left(x_1 - \bar{x}\right) = b_0^* + b_1 \cdot \left(x_i - \bar{x}\right).$$

- Potom můžeme psát:

$$\varphi = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left[Y_i - b_0^* - b_1 \cdot \left(x_i - \bar{x}\right)\right]^2.$$

Regresní a korelační analýza

- Jelikož hledáme minimum funkce φ s proměnnými b_0^* a b_1, položíme parciální derivace funkce φ rovny nule.

$$\frac{d\varphi}{db_0^*} = (-2)\cdot\sum_{i=1}^{n}\left[Y_i - b_0^* - b_1\cdot(x_i - \bar{x})\right] = 0$$

$$\frac{d\varphi}{db_1} = (-2)\cdot\sum_{i=1}^{n}\left[Y_i - b_0^* - b_1\cdot(x_i - \bar{x})\right]\cdot(x_i - \bar{x}) = 0$$

Regresní a korelační analýza

- Vyřešme nyní první rovnici.

$$(-2)\cdot\sum_{i=1}^{n}\left[Y_i - b_0^* - b_1\cdot(x_i-\bar{x})\right]=0$$

$$\sum_{i=1}^{n}Y_i - nb_0^* - b_1\cdot\sum_{i=1}^{n}(x_i-\bar{x})=0$$

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

$$nb_0^* = \sum_{i=1}^{n}Y_i$$

$$b_0^* = \frac{\sum_{i=1}^{n}Y_i}{n}$$

$$b_0^* = \bar{Y} \Rightarrow b_0 + b_1\bar{x} = \bar{Y}$$

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

Regresní a korelační analýza

- Nyní upravme druhou rovnici.

$$(-2)\cdot\sum_{i=1}^{n}\left[Y_i - b_0^* - b_1\cdot(x_i-\bar{x})\right]\cdot(x_i-\bar{x}) = 0$$

$$\sum_{i=1}^{n} Y_i\cdot(x_i-\bar{x}) - b_0^*\cdot\sum_{i=1}^{n}(x_i-\bar{x}) - b_1\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$$

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

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

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

Regresní a korelační analýza

- Odvodili jsme tedy vztahy pro koeficienty vyrovnávací přímky ve tvaru:

$$b_1 = \frac{\sum_{i=1}^{n} Y_i \cdot (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ a } b_0 = \bar{Y} - b_1 \cdot \bar{x} \ .$$

- Vyrovnávací přímka je potom ve tvaru:

$$\hat{Y}_i = b_0 + b_1 \cdot x_i = \bar{Y} - b_1 \cdot \bar{x} + b_1 \cdot x_i = \bar{Y} + b_1 \cdot (x_i - \bar{x}),$$

prochází tedy vždy bodem $[\bar{x}; \bar{Y}]$.

Regresní a korelační analýza

x_i	Y_i	$x_i - x_p$	$(x_i - x_p)\cdot Yi$	$(x_i - x_p)^2$
43	160	-148,25	-23720,00	21978,06
55	162	-136,25	-22072,50	18564,06
55	175	-136,25	-23843,75	18564,06
58	165	-133,25	-21986,25	17755,56
65	182	-126,25	-22977,50	15939,06
70	175	-121,25	-21218,75	14701,56
72	185	-119,25	-22061,25	14220,56
79	170	-112,25	-19082,50	12600,06
80	187	-111,25	-20803,75	12376,56
81	190	-110,25	-20947,50	12155,06
103	190	-88,25	-16767,50	7788,06
120	210	-71,25	-14962,50	5076,56
150	242	-41,25	-9982,50	1701,56
155	230	-36,25	-8337,50	1314,06
155	230	-36,25	-8337,50	1314,06
196	244	4,75	1159,00	22,56
350	306	158,75	48577,50	25201,56
448	322	256,75	82673,50	65920,56
736	407	544,75	221713,25	296752,56
754	342	562,75	192460,50	316687,56
x_p	Y_p		Σ	Σ
191,25	223,70		269482,50	880633,75

b_1	b_0
0,306	165,176

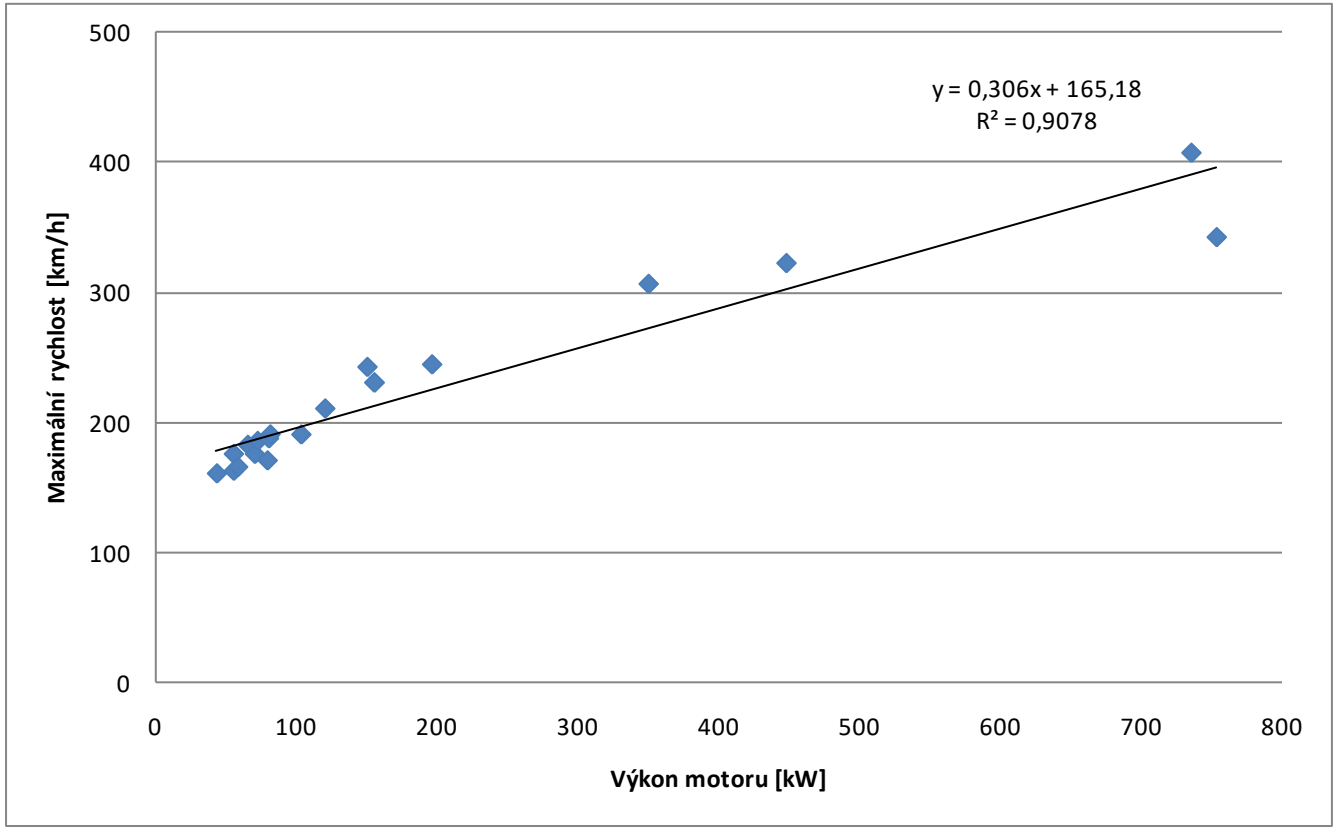
$$\hat{Y}_i = 165{,}176 + 0{,}306 x_i$$

Pozn.

$$x_p = \bar{x}$$

$$Y_p = \bar{Y}$$

Regresní a korelační analýza



$$\hat{Y}_i = 165{,}176 + 0{,}306 x_i$$

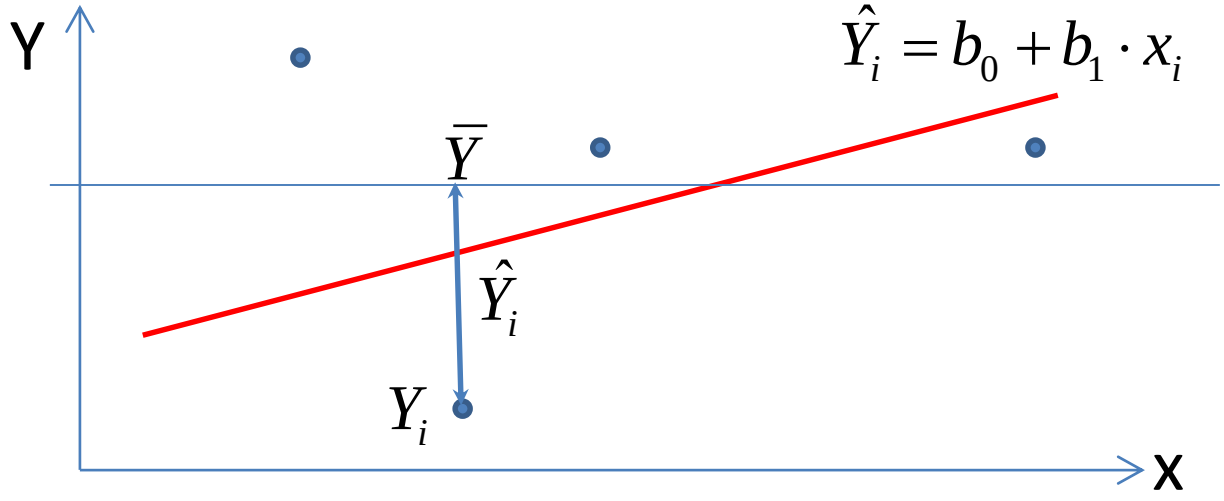
Regresní a korelační analýza

- Pro účely ověření správnosti zvoleného regresního modelu slouží **index determinace**.

$$SS_Y = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

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

$$SS_R = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$



Regresní a korelační analýza

- Označme:
  - Celkový součet čtverců $SS_Y = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$,
  - Součet čtverců modelu $SS_{\hat{Y}} = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$,
  - Reziduální součet čtverců $SS_R = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$.
- Platí:

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

Regresní a korelační analýza

- Zaveďme $\frac{SS_{\hat{Y}}}{SS_Y} + \frac{SS_R}{SS_Y} = 1$. Je zřejmé, že čím „lepší“ model bude, tím více se bude první zlomek blížit k 1 a naopak.

- Zaveďme index determinace $R^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$.
  - Index determinace nabývá hodnot z intervalu $\langle 0;1 \rangle$. Velké hodnoty (cca nad 0,8) znamenají, že použitý regresní model se hodí pro popis závislosti.

Regresní a korelační analýza

$\hat{Y}_i$	$(\hat{Y}_i - Y_p)^2$	$(Y_i - Y_p)^2$
178,33	2058,07	4057,69
182,01	1738,38	3806,89
182,01	1738,38	2371,69
182,92	1662,67	3445,69
185,07	1492,57	1738,89
186,60	1376,68	2371,69
187,21	1331,64	1497,69
189,35	1179,89	2883,69
189,66	1158,97	1346,89
189,96	1138,22	1135,69
196,69	729,29	1135,69
201,90	475,38	187,69
211,08	159,34	334,89
212,61	123,05	39,69
212,61	123,05	39,69
225,15	2,11	412,09
272,28	2359,92	6773,29
302,27	6172,93	9662,89
390,40	27788,49	33598,89
395,91	29655,24	13994,89
	Σ	Σ
	82464,27	90836,20

R^2
0,908

Regresní a korelační analýza

- Odhad regresní funkce nám umožňuje predikovat hodnotu Y_0 při libovolné hodnotě x_0:
  - Je-li $x_0 \in \langle x_1; x_n \rangle$, potom hovoříme o **interpolaci**.
  - Je-li $x_0 \notin \langle x_1; x_n \rangle$, potom se jedná o **extrapolaci**.

Regresní a korelační analýza

- My jsme se zatím zabývali pouze případem, kdy vyrovnávací křivkou byla přímka. V praxi se používají i jiné regresní modely:
  1. Parabolická regrese:
     $$EY_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 .$$
  2. Polynomická regrese *n*-tého stupně:
     $$EY_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \ldots + \beta_n \cdot x_i^n .$$
  3. Hyperbolická regrese:
     $$EY_i = \beta_0 + \frac{\beta_1}{x_i} .$$

Regresní a korelační analýza

4. Logaritmická regrese:

$$EY_i = \beta_0 + \beta_1 \cdot \log x_i .$$

5. Exponenciální regrese:

$$EY_i = \beta_0 \cdot \beta_1^{x_i} .$$

Regresní a korelační analýza

1) Uvažujme parabolickou regresi, vyrovnávací křivka (její odhad) je tedy vyjádřena ve tvaru:

$$\widehat{Y}_i = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2 .$$

Jelikož se jedná o regresní model lineární v parametrech, můžeme pro odhad koeficientů použít metodu nejmenších čtverců.

Regresní a korelační analýza

$$\text{Tedy } \varphi = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2\right)^2 \to \min.$$

Hledáme minimum, položíme parciální derivace rovny nule :

$$\frac{d\varphi}{db_0} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2\right)^1 \cdot (-1) = 0,$$

$$\frac{d\varphi}{db_1} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2\right)^1 \cdot (-x_i) = 0,$$

$$\frac{d\varphi}{db_2} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2\right)^1 \cdot (-x_i^2) = 0.$$

Regresní a korelační analýza

Získanou soustavu upravíme:

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 \cdot x_i - \sum_{i=1}^{n} b_2 \cdot x_i^2 = 0,$$

$$\sum_{i=1}^{n} Y_i \cdot x_i - \sum_{i=1}^{n} b_0 \cdot x_i - \sum_{i=1}^{n} b_1 \cdot x_i^2 - \sum_{i=1}^{n} b_2 \cdot x_i^3 = 0,$$

$$\sum_{i=1}^{n} Y_i \cdot x_i^2 - \sum_{i=1}^{n} b_0 \cdot x_i^2 - \sum_{i=1}^{n} b_1 \cdot x_i^3 - \sum_{i=1}^{n} b_2 \cdot x_i^4 = 0.$$

Regresní a korelační analýza

Dalšími úpravami dostaneme :

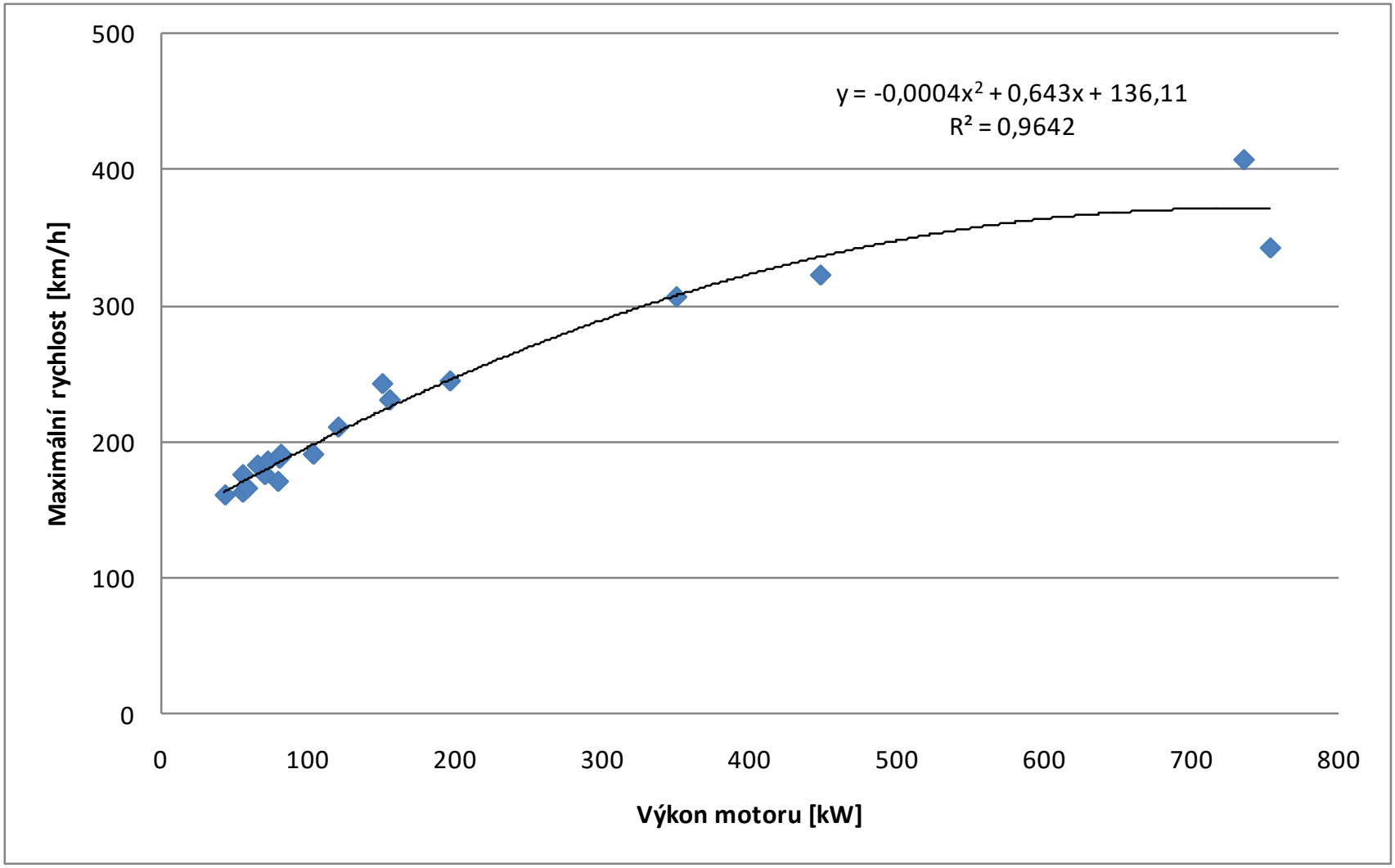
$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2,$$

$$\sum_{i=1}^{n} Y_i \cdot x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3,$$

$$\sum_{i=1}^{n} Y_i \cdot x_i^2 = b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4.$$

Získali jsme soustavu tří rovnic se třemi neznámými, řešením získáme odhady koeficientů regresního modelu.

Regresní a korelační analýza



Regresní a korelační analýza

2) Uvažujme polynomickou regresi, vyrovnávací křivka (její odhad) je vyjádřena ve tvaru:

$$\widehat{Y}_i = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2 + \ldots + b_n \cdot x_i^n .$$

Jelikož se opět jedná o regresní model lineární v parametrech, můžeme pro odhad koeficientů použít metodu nejmenších čtverců.

Regresní a korelační analýza

$$\text{Tedy } \varphi = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2 - \ldots - b_n \cdot x_i^n\right)^2 \rightarrow \min.$$

Hledáme minimum, položíme parciální derivace rovny nule :

$$\frac{d\varphi}{db_0} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2 - \ldots - b_n \cdot x_i^n\right)^1 \cdot (-1) = 0,$$

$$\frac{d\varphi}{db_1} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2 - \ldots - b_n \cdot x_i^n\right)^1 \cdot (-x_i) = 0,$$

$$\frac{d\varphi}{db_2} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2 - \ldots - b_n \cdot x_i^n\right)^1 \cdot \left(-x_i^2\right) = 0,$$

$$\vdots$$

$$\frac{d\varphi}{db_n} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_i - b_2 \cdot x_i^2 - \ldots - b_n \cdot x_i^n\right)^1 \cdot \left(-x_i^n\right) = 0.$$

Regresní a korelační analýza

Po úpravách dostaneme:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 + \ldots + b_n \sum_{i=1}^{n} x_i^n,$$

$$\sum_{i=1}^{n} Y_i \cdot x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 + \ldots + b_n \sum_{i=1}^{n} x_i^{n+1},$$

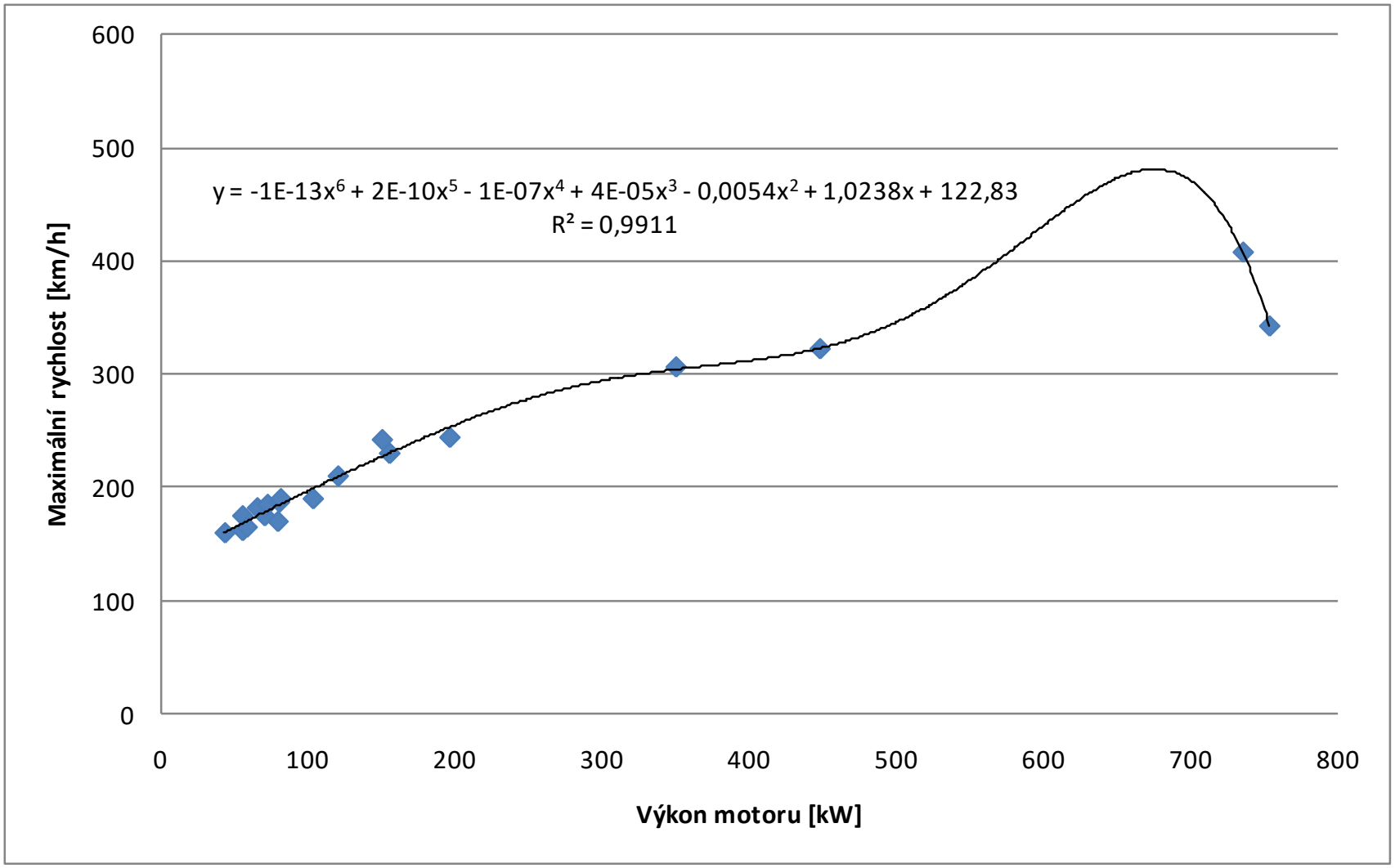
$$\sum_{i=1}^{n} Y_i \cdot x_i^2 = b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 + \ldots + b_n \sum_{i=1}^{n} x_i^{n+2},$$

$$\vdots$$

$$\sum_{i=1}^{n} Y_i \cdot x_i^n = b_0 \sum_{i=1}^{n} x_i^n + b_1 \sum_{i=1}^{n} x_i^{n+1} + b_2 \sum_{i=1}^{n} x_i^{n+2} + \ldots + b_n \sum_{i=1}^{n} x_i^{2n}.$$

Získali jsme soustavu (n+1) rovnic s (n+1) neznámými, řešením získáme odhady koeficientů regresního modelu.

Regresní a korelační analýza



Regresní a korelační analýza

3) Uvažujme hyperbolickou regresi, vyrovnávací křivka (její odhad) je vyjádřena ve tvaru:

$$\widehat{Y}_i = b_0 + \frac{b_1}{x_i}.$$

Jelikož se opět jedná o regresní model lineární v parametrech, můžeme pro odhad koeficientů použít metodu nejmenších čtverců.

Regresní a korelační analýza

$$\text{Tedy } \varphi = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - \frac{b_1}{x_i} \right)^2 \rightarrow \min .$$

Hledáme minimum, položíme parciální derivace rovny nule :

$$\frac{d\varphi}{db_0} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - \frac{b_1}{x_i} \right)^1 \cdot (-1) = 0,$$

$$\frac{d\varphi}{db_1} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - \frac{b_1}{x_i} \right)^1 \cdot \left(-\frac{1}{x_i} \right) = 0.$$

Regresní a korelační analýza

Rovnice upravíme:

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} \frac{b_1}{x_i} = 0,$$

$$\sum_{i=1}^{n} \frac{Y_i}{x_i} - \sum_{i=1}^{n} \frac{b_0}{x_i} - \sum_{i=1}^{n} \frac{b_1}{x_i^2} = 0,$$

– – – – – – – – – – – – – – – – –

$$1) \sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} \frac{1}{x_i} = 0,$$

$$2) \sum_{i=1}^{n} \frac{Y_i}{x_i} - b_0 \sum_{i=1}^{n} \frac{1}{x_i} - b_1 \sum_{i=1}^{n} \frac{1}{x_i^2} = 0.$$

Regresní a korelační analýza

Postupně vyjádříme:

$$\text{z } 1)\; b_0 = \frac{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} \frac{1}{x_i}}{n} \text{ a } b_1 = \frac{\sum_{i=1}^{n} Y_i - n b_0}{\sum_{i=1}^{n} \frac{1}{x_i}},$$

$$\text{z } 2)\; b_0 = \frac{\sum_{i=1}^{n} \frac{Y_i}{x_i} - b_1 \sum_{i=1}^{n} \frac{1}{x_i^2}}{\sum_{i=1}^{n} \frac{1}{x_i}} \text{ a } b_1 = \frac{\sum_{i=1}^{n} \frac{Y_i}{x_i} - b_0 \sum_{i=1}^{n} \frac{1}{x_i}}{\sum_{i=1}^{n} \frac{1}{x_i^2}}.$$

Regresní a korelační analýza

Jelikož musí platit :

$$\frac{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} \frac{1}{x_i}}{n} = \frac{\sum_{i=1}^{n} \frac{Y_i}{x_i} - b_1 \sum_{i=1}^{n} \frac{1}{x_i^2}}{\sum_{i=1}^{n} \frac{1}{x_i}},$$

dostaneme po úpravách :

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

Také platí :

$$\frac{\sum_{i=1}^{n} Y_i - nb_0}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{\sum_{i=1}^{n} \frac{Y_i}{x_i} - b_0 \sum_{i=1}^{n} \frac{1}{x_i}}{\sum_{i=1}^{n} \frac{1}{x_i^2}},$$

potom :

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

Regresní a korelační analýza

4) Uvažujme logaritmickou regresi, vyrovnávací křivka (její odhad) je vyjádřena ve tvaru:

$$\widehat{Y}_i = b_0 + b_1 \cdot \log x_i .$$

Jelikož se opět jedná o regresní model lineární v parametrech, můžeme pro odhad koeficientů použít metodu nejmenších čtverců.

Regresní a korelační analýza

$$\text{Tedy } \varphi = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot \log x_i\right)^2 \to \min.$$

Hledáme minimum, položíme parciální derivace rovny nule :

$$\frac{d\varphi}{db_0} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot \log x_i\right)^1 \cdot \left(-1\right) = 0,$$

$$\frac{d\varphi}{db_1} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot \log x_i\right)^1 \cdot \left(-\log x_i\right) = 0.$$

Regresní a korelační analýza

Úpravami dostaneme:

$$1)\ \sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} \log x_i = 0,$$

$$2)\ \sum_{i=1}^{n} Y_i \cdot \log x_i - b_0 \sum_{i=1}^{n} \log x_i - b_1 \sum_{i=1}^{n} \log^2 x_i = 0.$$

Regresní a korelační analýza

Postupně vyjádříme:

$$\text{z } 1)\; b_0 = \frac{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} \log x_i}{n} \text{ a } b_1 = \frac{\sum_{i=1}^{n} Y_i - n b_0}{\sum_{i=1}^{n} \log x_i},$$

$$\text{z } 2)\; b_0 = \frac{\sum_{i=1}^{n} Y_i \cdot \log x_i - b_1 \sum_{i=1}^{n} \log^2 x_i}{\sum_{i=1}^{n} \log x_i} \text{ a } b_1 = \frac{\sum_{i=1}^{n} Y_i \cdot \log x_i - b_0 \sum_{i=1}^{n} \log x_i}{\sum_{i=1}^{n} \log^2 x_i}.$$

Regresní a korelační analýza

Jelikož platí:

$$\frac{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} \log x_i}{n} = \frac{\sum_{i=1}^{n} Y_i \cdot \log x_i - b_1 \sum_{i=1}^{n} \log^2 x_i}{\sum_{i=1}^{n} \log x_i},$$

získáme:

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

Regresní a korelační analýza

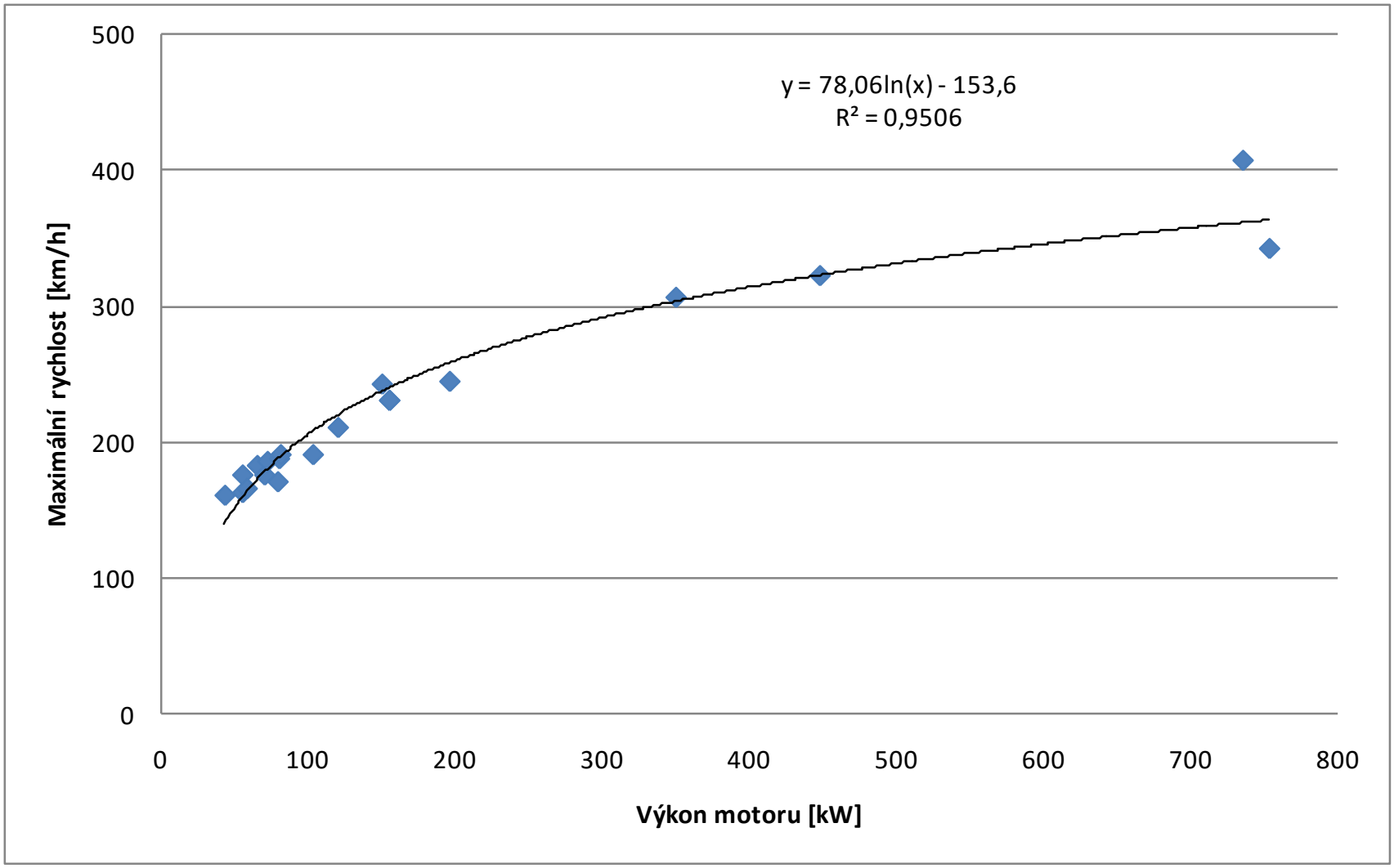
Dále platí :

$$\frac{\sum_{i=1}^{n} Y_i - nb_0}{\sum_{i=1}^{n} \log x_i} = \frac{\sum_{i=1}^{n} Y_i \cdot \log x_i - b_0 \sum_{i=1}^{n} \log x_i}{\sum_{i=1}^{n} \log^2 x_i},$$

získáme :

$$b_0 = \frac{\sum_{i=1}^{n} Y_i \cdot \log^2 x_i - \sum_{i=1}^{n} Y_i \cdot \log x_i \cdot \sum_{i=1}^{n} \log x_i}{n\sum_{i=1}^{n} \log^2 x_i - \left(\sum_{i=1}^{n} \log x_i\right)^2}.$$

Regresní a korelační analýza



Regresní a korelační analýza

5) Uvažujme exponenciální regresi, vyrovnávací křivka (její odhad) je vyjádřena ve tvaru:

$$\widehat{Y}_i = b_0 \cdot b_1^{x_i}.$$

Tento model není lineární v parametrech, použití metody nejmenších čtverců je problematické, výstupem jsou nelineární rovnice. V tomto případě užijeme **linearizující transformaci**.

Regresní a korelační analýza

Postupně upravíme:

$$\widehat{Y}_i = b_0 \cdot b_1^{x_i} \ / \log,$$

$$\log \widehat{Y}_i = \log\left(b_0 \cdot b_1^{x_i}\right),$$

$$\log \widehat{Y}_i = \log b_0 + x_i \cdot \log b_1.$$

Pokud $A = \log b_0, B = \log b_1$, potom lze psát:

$$\log \widehat{Y}_i = A + B \cdot x_i.$$

Nyní již můžeme použít metodu nejmenších čtverců, ale v logaritmickém tvaru:

$$\varphi = \sum_{i=1}^{n} \left(\log Y_i - \log \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(\log Y_i - A - Bx_i\right)^2 \rightarrow \min.$$

Regresní a korelační analýza

Hledáme minimum, položíme parciální derivace rovny nule :

$$\frac{d\varphi}{dA} = 2\sum_{i=1}^{n}\left(\log Y_i - A - Bx_i\right)^1 \cdot \left(-1\right) = 0,$$

$$\frac{d\varphi}{dB} = 2\sum_{i=1}^{n}\left(\log Y_i - A - Bx_i\right)^1 \cdot \left(-x_i\right) = 0.$$

Upravíme :

$$1)\ \sum_{i=1}^{n}\log Y_i - nA - B\sum_{i=1}^{n}x_i = 0,$$

$$2)\ \sum_{i=1}^{n}x_i \cdot \log Y_i - A\sum_{i=1}^{n}x_i - B\sum_{i=1}^{n}x_i^2 = 0.$$

Regresní a korelační analýza

Postupně vyjádříme:

$$\text{z } 1)\ A = \frac{\sum_{i=1}^{n} \log Y_i - B\sum_{i=1}^{n} x_i}{n} \text{ a } B = \frac{\sum_{i=1}^{n} \log Y_i - nA}{\sum_{i=1}^{n} x_i},$$

$$\text{z } 2)\ A = \frac{\sum_{i=1}^{n} x_i \cdot \log Y_i - B\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} \text{ a } B = \frac{\sum_{i=1}^{n} x_i \cdot \log Y_i - A\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}.$$

Regresní a korelační analýza

Jelikož musí platit :

$$\frac{\sum_{i=1}^{n} \log Y_i - B\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} x_i \cdot \log Y_i - B\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i},$$

dostaneme :

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

Regresní a korelační analýza

Dále musí platit :

$$\frac{\sum_{i=1}^{n} \log Y_i - nA}{\sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{n} x_i \cdot \log Y_i - A\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2},$$

tedy :

$$A = \frac{\sum_{i=1}^{n} \log Y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \cdot \log Y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \log b_0.$$

Regresní a korelační analýza

- Jelikož jsme použili metodu nejmenších čtverců v logaritmické formě, je nutno přistoupit ke stanovení indexu determinace rovněž v logaritmické formě:

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

Regresní a korelační analýza

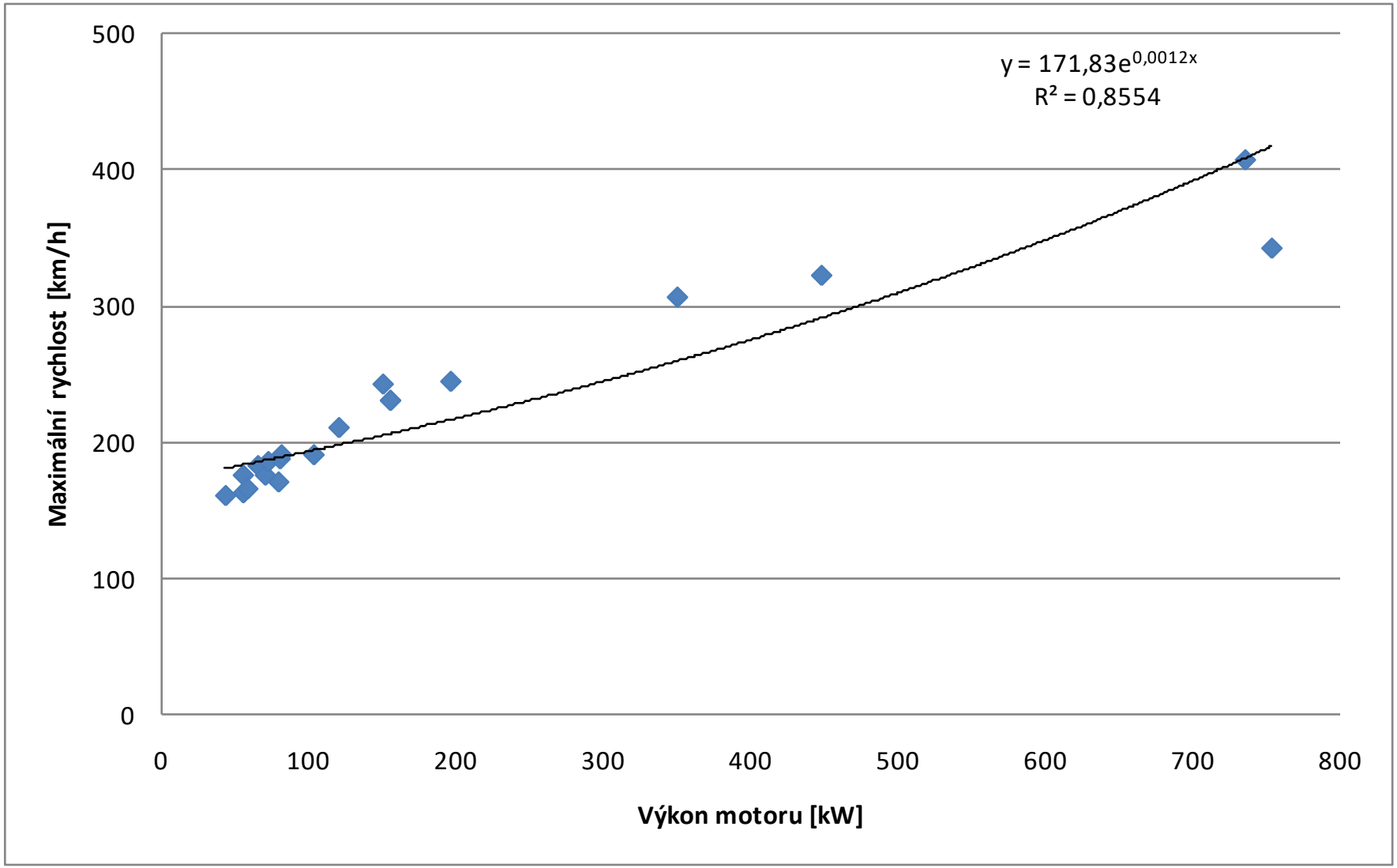
- Excel používá při exponenciální regresi jiný zápis regresní funkce:

  $\widehat{Y}_i = b_0 \cdot e^{b_1' x_i}$.

  Označme $b_1 = e^{b_1'}$. Jelikož platí $a^y = x \Rightarrow \log_a x = y$, můžeme psát:

  $b_1' = \ln b_1$.

Regresní a korelační analýza



Regresní a korelační analýza

- Doposud jsme se zabývali vystižením závislosti vysvětlované proměnné na jedné vysvětlující proměnné, tedy jednoduchou regresí.
- Podívejme se nyní na vícenásobnou regresi, vysvětlovaná proměnná Y_i závisí na několika vysvětlujících proměnných $x_{1i}, x_{2i}, \ldots, x_{ni}$.
- Pro jednoduchost se zaměřme pouze na závislost na dvou vysvětlujících proměnných.

Regresní a korelační analýza

- Odhad regresní funkce můžeme zapsat ve tvaru:

$$\widehat{Y_i} = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} ,$$

kde parametry b_1 a b_2 se nazývají dílčí regresní koeficienty a udávají, jak se průměrně změní vysvětlovaná proměnná při jednotkové změně příslušné vysvětlující proměnné.

Regresní a korelační analýza

- Jelikož se jedná o model lineární v parametrech, lze použít metodu nejmenších čtverců, tedy:

$$\varphi = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_{1i} - b_2 \cdot x_{2i}\right)^2 \rightarrow \min,$$

$$\frac{d\varphi}{db_0} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_{1i} - b_2 \cdot x_{2i}\right)^1 \cdot \left(-1\right) = 0,$$

$$\frac{d\varphi}{db_1} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_{1i} - b_2 \cdot x_{2i}\right)^1 \cdot \left(-x_{1i}\right) = 0,$$

$$\frac{d\varphi}{db_2} = 2\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 \cdot x_{1i} - b_2 \cdot x_{2i}\right)^1 \cdot \left(-x_{2i}\right) = 0.$$

Regresní a korelační analýza

- Po úpravách získáme:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i},$$

$$\sum_{i=1}^{n} Y_i \cdot x_{1i} = b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i} \cdot x_{2i},$$

$$\sum_{i=1}^{n} Y_i \cdot x_{2i} = b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i} \cdot x_{2i} + b_2 \sum_{i=1}^{n} x_{21i}^2.$$

- Řešením této soustavy získáme odhady regresních koeficientů.

Regresní a korelační analýza

- Pro posouzení a srovnání individuálního vlivu jednotlivých vysvětlujících proměnných na vysvětlovanou proměnnou zavádíme normalizované regresní koeficienty – **B-koeficienty**:

$$B_1 = \frac{s_{x_1}}{s_Y} \cdot b_1, B_2 = \frac{s_{x2}}{s_Y} \cdot b_2 ,$$

kde s_{x_1}, s_{x_2} a s_Y jsou výběrové směrodatné odchylky jednotlivých proměnných.

Regresní a korelační analýza

- Známe-li jednoduché korelační koeficienty, můžeme psát:

$$B_1 = \frac{r_{x_1,Y} - r_{x_2,Y} \cdot r_{x_1,x_2}}{1 - r_{x_1,x_2}^2}, B_2 = \frac{r_{x_2,Y} - r_{x_1,Y} \cdot r_{x_1,x_2}}{1 - r_{x_1,x_2}^2}.$$

- B-koeficienty zavádíme, abychom mohli srovnat intenzity vlivu jednotlivých vysvětlujících proměnných na vysvětlovanou proměnnou.

Regresní a korelační analýza

- **Př.** Uvažujme závislost maximální rychlosti osobního automobilu v [km/h] na výkonu motoru [kW] a točivém momentu [Nm]. Výpočtem jsme zjistili dílčí regresní koeficienty:

$$b_1 = 0{,}78\,\frac{\text{km/h}}{\text{kW}} \text{ a } b_2 = 0{,}47\,\frac{\text{km/h}}{\text{Nm}}\,.$$

- Zajímá nás, vliv které vysvětlující proměnné je větší.

Regresní a korelační analýza

- Dílčí koeficienty nelze přímo srovnat, protože jsou v jiných jednotkách. Proto je pro srovnání nutno provést výpočet B-koeficientů.
- Uvažujme, že známe výběrové směrodatné odchylky jednotlivých proměnných, tedy:

$$s_Y = 87{,}24\,\text{km/h},\ s_{x_1} = 101{,}28\,\text{kW},\ s_{x_2} = 169{,}29\,\text{Nm}.$$

Regresní a korelační analýza

- Dosazením a výpočtem dostaneme:

$$B_1 = \frac{s_{x_1}}{x_Y} \cdot b_1 = \frac{101{,}28}{87{,}24} \cdot 0{,}78 \doteq 0{,}90,$$

$$B_2 = \frac{s_{x2}}{x_Y} \cdot b_2 = \frac{169{,}29}{87{,}24} \cdot 0{,}47 \doteq 0{,}91.$$

- Z výsledků vidíme, že vliv obou vysvětlujících proměnných na maximální rychlost je zhruba stejný.

Regresní a korelační analýza

- Pro stanovení síly závislostí užíváme **koeficienty dílčí korelace** nebo **koeficienty vícenásobné korelace**.
- Koeficienty dílčí korelace vyjadřují sílu závislosti mezi vysvětlovanou proměnnou a příslušnou vysvětlující proměnnou oproštěnou od vlivu druhé vysvětlující proměnné.

Regresní a korelační analýza

- Příslušné dílčí korelační koeficienty stanovíme dle vztahů:

$$r_{x_1,Y\bullet x_2} = \frac{r_{x_1,Y} - r_{x_2,Y} \cdot r_{x_1,x_2}}{\sqrt{\left(1 - r_{x_2,Y}^2\right) \cdot \left(1 - r_{x_1,x_2}^2\right)}},$$

$$r_{x_2,Y\bullet x_1} = \frac{r_{x_2,Y} - r_{x_1,Y} \cdot r_{x_1,x_2}}{\sqrt{\left(1 - r_{x_1,Y}^2\right) \cdot \left(1 - r_{x_1,x_2}^2\right)}}.$$

Regresní a korelační analýza

- Koeficient vícenásobné korelace vyjadřuje sílu závislosti vysvětlované proměnné na všech vysvětlujících proměnných. Určíme ho podle vztahu:

$$r_{x_1,x_2,Y} = \sqrt{\frac{r_{x_1,Y}^2 - 2r_{x_1,Y} \cdot r_{x_2,Y} \cdot r_{x_1,x_2} + r_{x_2,Y}^2}{1 - r_{x_1,x_2}^2}}.$$