

## Nové možnosti rozvoje vzdělávání na Technické univerzitě v Liberci

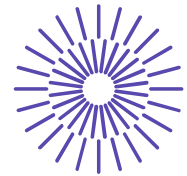
Specifický cíl A3: Tvorba nových profesně zaměřených studijních programů

NPO\_TUL\_MSMT-16598/2022



## Analýza závislostí – část 3

Ing. Vladimíra Hovorková Valentová, Ph.D.



## Analýza závislostí číselných proměnných (regresní a korelační analýza)

- zkoumání závislosti dvou event. více proměnných, měření síly této závislosti, atd.
- cílem je hlubší vniknutí do podstaty sledovaných jevů a procesů, přiblížení k tzv. příčinným souvislostem.

### Korelační tabulka

- dvourozměrná tabulka, ve které jsou uspořádány numerické proměnné.

#### Korelační tabulka

$x_i \backslash y_j$	$y_1$	$y_2$	$\dots$	$y_l$	Součty četností $n_{i\bullet}$
$x_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1l}$	$n_{1\bullet}$
$x_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2l}$	$n_{2\bullet}$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$x_k$	$n_{k1}$	$n_{k2}$	$\dots$	$n_{kl}$	$n_{k\bullet}$
Součty četností $n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet l}$	$n$

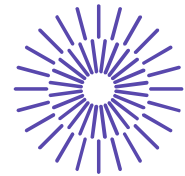
#### Symbolika:

$n_{ij}$  ..... sdružené (simultánní) absolutní četnosti

$n_{i\bullet}$ ,  $n_{\bullet j}$  ..... okrajové (marginální) absolutní četnosti

$$n_{i\bullet} = \sum_{j=1}^l n_{ij}; \quad n_{\bullet j} = \sum_{i=1}^k n_{ij}$$

$$\sum_{i=1}^k n_{i\bullet} = \sum_{j=1}^l n_{\bullet j} = \sum_{i=1}^k \sum_{j=1}^l n_{ij} = n$$



### Podmíněné rozdělení četností:

Rozdělení četností jedné proměnné, které odpovídá určité obměně druhé proměnné (tj. za podmínky, že druhá proměnná nabyla určité obměny).

$$\text{Podmíněný průměr: } \bar{y}_i = \frac{\sum_{j=1}^l y_j n_{ij}}{n_{i\bullet}}$$

$$\text{Podmíněný rozptyl: } s_i^2 = \frac{\sum_{j=1}^l (y_j - \bar{y}_i)^2 n_{ij}}{n_{i\bullet}}$$

## Grafické znázornění dvourozměrného rozdělení četností

- je další formou popisu závislosti;

- různé typy grafů:

- ❖ čára podmíněných průměrů,
- ❖ čára podmíněných rozptylů,
- ❖ bodový graf (diagram).

## Regresní analýza

- zkoumání jednostranné závislosti proměnné  $y$  (závislá, vysvětlovaná) na proměnné  $x$  (nezávislá, vysvětlující), resp. na proměnných  $x_1, x_2, \dots, x_k$ ;

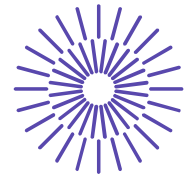
- nezávislá proměnná = příčina, závislá proměnná = důsledek;

- důležitý je přitom směr závislosti, tzn. která proměnná je závislá, a která nezávislá;

- závislost většinou modelujeme nějakou matematickou funkcí (tzv. regresní funkce).

### Postup regresní analýzy:

1. Volba typu regresní funkce.
2. Odhad parametrů zvolené regresní funkce.
3. Testování hypotéz o parametrech regresní funkce.
4. Ověření vhodnosti zvoleného regresního modelu.



## Jednoduchá regresní analýza

### 1. Volba typu regresní funkce

- při volbě typu regresní funkce lze uplatnit různá kritéria;
- volba by se měla v první řadě opírat o určitou teorii, tzn. vyplývat z věcného rozboru vztahů proměnných;
- vždy se snažíme o jednoduchost modelu (ne příliš mnoho parametrů);
- je třeba změřit vhodnou charakteristikou přilnavost regresní funkce k datům.

### 2. Odhad parametrů regresní funkce

#### Regresní modely

- matematické modely, které vyjadřují představu o průběhu závislosti proměnných;
- umožňují odhady neznámých hodnot závisle proměnné  $y$  ze známých hodnot nezávisle proměnné  $x$ , resp. nezávisle proměnných  $x_1, x_2, \dots, x_k$ .

#### Obecný tvar modelu:

$$y_i = \eta_i + \varepsilon_i = \eta(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

**Symbolika:**  $\eta_i$  ... deterministická složka

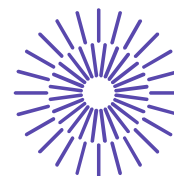
$\varepsilon_i$  ... náhodná (rušivá) složka

**Typy modelů:** 1. **aditivní (součtový)** – jeho složky se skládají sčítáním, je nejčastější.

2. **multiplikativní (součinnový)** – jeho složky se skládají násobením.

**Teoretická regresní funkce:**  $\eta = \eta(x)$

- existují různé typy regresních funkcí;
- nejčastější jsou lineární regresní funkce;
- linearita se může hodnotit jak z hlediska proměnných, tak z hlediska parametrů;



- každá regresní funkce má určitý počet parametrů (jejich počet je  $p$ ).

**Parametry regresní funkce:**

- neznámé konstanty; symbolicky je značíme řeckými písmeny  $(\beta_0, \beta_1, \dots, \beta_k)$ ;

- jejich hodnoty lze odhadnout z výběrových dat;

- k jejich odhadu je třeba zvolit takovou metodu, aby odhady měly co nejlepší vlastnosti.

**1) Funkce lineární z hlediska parametrů**

**přímka**  $\eta = \beta_0 + \beta_1 x$

**rovina**  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

**nadrovina**  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$

**parabola**  $\eta = \beta_0 + \beta_1 x + \beta_2 x^2$

**hyperbola**  $\eta = \beta_0 + \beta_1 x^{-1}$

**logaritmická funkce**  $\eta = \beta_0 + \beta_1 \ln x$

**polynom**  $\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k$

**2) Funkce nelineární z hlediska parametrů**

**exponenciální funkce**  $\eta = \beta_0 \beta_1^x$

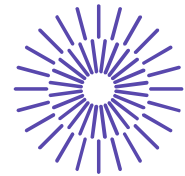
**mocninná funkce**  $\eta = \beta_0 x^{\beta_1}$

**Törnquistova křivka**  $\eta = \frac{\beta_0 x}{x + \beta_1}$

**I. Jednoduchá lineární regrese**

- regresní funkce je lineární z hlediska parametrů;

- má jednu vysvětlující proměnnou (regresor)  $x$ .



**Teoretická (hypotetická) regresní funkce:**  $\eta = \beta_0 + \beta_1 x$

-  $\beta_0, \beta_1 \dots$  parametry;  $x \dots$  regresor

- nutno provést odhad teoretické regresní funkce, tzn. odhad neznámých parametrů  $\beta_0, \beta_1$ ;

- nejlepší metodou odhadu parametrů lineární regresní funkce je **metoda nejmenších čtverců**;

- takový postup zaručí, že výběrová regresní funkce bude co nejlépe přiléhat k výběrovým hodnotám.

**Empirická (výběrová) regresní funkce:**  $\hat{\eta} = Y = b_0 + b_1 x$

$b_0, b_1 \dots$  odhady parametrů;  $b_0 = \hat{\beta}_0$ ;  $b_1 = \hat{\beta}_1$

- odhad parametrů se provádí metodou nejmenších čtverců;

- když odhadneme parametry, získáme tzv. výběrovou regresní funkci.

### **Metoda nejmenších čtverců**

- lze ji použít pouze k odhadu parametrů funkcí lineárních v parametrech (v lineární regresi);

- princip: parametry odhadujeme tak, aby pro ně byl minimální součet čtverců reziduí.

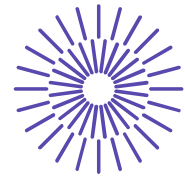
$$y_i = \eta_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$$y_i = Y_i + \hat{\varepsilon}_i = b_0 + b_1 x_i + \hat{\varepsilon}_i$$

**Reziduum:**  $\hat{\varepsilon}_i = y_i - Y_i = y_i - b_0 - b_1 x_i = e_i$

$$S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \quad \Rightarrow \quad \text{minimalizovat}$$

1. Stanovíme parciální derivace a položíme je rovny 0.
2. Vznikne soustava dvou rovnic (tzv. normální rovnice).
3. Vyřešíme je a získáme vzorce pro výpočet  $b_0$  a  $b_1$ .



**Vzorce pro výpočet parametrů výběrové regresní přímky:**

$$b_1 = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \frac{\sum_{i=1}^n x_i}{n} \cdot \frac{\sum_{i=1}^n y_i}{n}}{\frac{\sum_{i=1}^n x_i^2}{n} - \left( \frac{\sum_{i=1}^n x_i}{n} \right)^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x^2} = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \frac{\sum y_i \cdot \sum x_i^2 - \sum x_i \cdot \sum x_i \cdot y_i}{n \sum x_i^2 - (\sum x_i)^2} = \bar{y} - b_1 \cdot \bar{x}$$

$b_1$  ... **výběrový regresní koeficient (směrnice výběrové regresní přímky)**

- udává, jak velká průměrná změna proměnné y odpovídá zvýšení proměnné x o jednotku.

$s_{xy}$  ... **kovariance**

- symetrická míra, tzn.  $s_{xy} = s_{yx}$ .

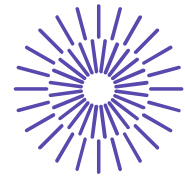
## **II. Nelineární regrese**

- není-li regresní funkce lineární v parametrech, nelze její parametry odhadnout metodou nejmenších čtverců;
- pro odhad parametrů se používá řada různých metod;
- častá je metoda linearizující transformace (např. zlogaritmování);
- většinou následují další metody pro zlepšení vlastností odhadů;
- výpočetně značně náročné (využití statistických programů).

## **3. Testování hypotéz o parametrech regresní funkce**

### **t – testy**

- dílčí testy o nulových hodnotách jednotlivých regresních parametrů.



$$1) H_0 : \beta_j = 0, j = 1, 2, \dots, k$$

$$H_1 : \text{non } H_0$$

**2) Testové kritérium:**

$$t = \frac{b_j}{s(b_j)}, \quad j = 1, 2, \dots, k. \quad \text{Statistika } t \text{ má při platnosti } H_0 \text{ rozdělení } t \text{ s } (n-p) \text{ stupni volnosti.}$$

**3) Kritický obor:**

$$W \equiv \left\{ t; t \leq t_{\frac{\alpha}{2}}(n-p) \cup t \geq t_{1-\frac{\alpha}{2}}(n-p) \right\}$$

**4) Závěr testu:**

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme  $H_0$  a přijímáme  $H_1$ , jinak řečeno, test je statisticky významný. Testovaný parametr je statisticky významný, je v regresní funkci přínosný.

## 4. Ověření vhodnosti zvoleného regresního modelu

### Celkový F – test

- testujeme vhodnost modelu jako celku;

- analýza rozptylu.

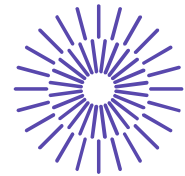
$$1) H_0 : \beta_0 = c, \beta_1, \beta_2, \dots, \beta_k = 0 \text{ (regresní funkce nemá žádný význam, tj. není vhodná)}$$

$$H_1 : \text{non } H_0$$

**2) Testové kritérium:**

$$F = \frac{S_T/p-1}{S_R/n-p}; \quad \text{Statistika } F \text{ má při platnosti } H_0 \text{ rozdělení } F \text{ s } (p-1) \text{ a } (n-p) \text{ stupni volnosti.}$$





**Rozklad celkového součtu čtverců:**  $S_y = S_T + S_R$

$S_y$  ... **celkový součet čtverců**; charakterizuje celkovou variabilitu proměnné  $y$ .

$S_T$  ... **teoretický součet čtverců**; část variability, kterou lze vysvětlit zvolenou regresní funkcí.

$S_R$  ... **reziduální součet čtverců**; část variability, kterou nelze zvolenou regresní funkcí vysvětlit.

$$S_y = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$S_T = \sum_{i=1}^n (Y_i - \bar{y})^2 ; \quad S_R = \sum_{i=1}^n (y_i - Y_i)^2 .$$

### 3) Kritický obor:

$$W \equiv \{F; F > F_{1-\alpha}(p-1; n-p)\}$$

### 4) Závěr testu:

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme  $H_0$  a přijímáme  $H_1$ , jinak řečeno, test je statisticky významný. Model lze považovat za vhodný.

## Kritéria pro posouzení kvality regresní funkce

### 1. Index determinace

- za vhodnější se považuje ta regresní funkce, u které je hodnota  $I^2$  vyšší.

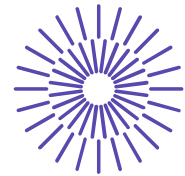
$$\text{Index determinace: } I^2 = \frac{S_T}{S_y}; \quad I^2 \in \langle 0; 1 \rangle$$

- udává, jaký podíl variability proměnné  $y$  lze vysvětlit zvolenou regresní funkcí (lze udávat i v %).

- je zároveň mírou těsnosti závislosti proměnné  $y$  na proměnné  $x$ .

- obecná míra, nezávislá na typu regresní funkce (lze použít i pro měření nelineární závislosti).

- tato míra není symetrická.



*Pozn.:* Při srovnávání funkcí s rozdílným počtem parametrů musíme hodnotu  $I^2$  upravit (penalizovat), neboť u funkce s vyšším počtem parametrů vychází hodnota  $I^2$  automaticky vyšší. Existují různé formy penalizace, např.:

$$I_{adj}^2 = 1 - (1 - I^2) \cdot \frac{n-1}{n-p} = 1 - \frac{(n-1)S_R}{(n-p)S_y}$$

*Pozn.:* *adjusted* = upravený.

**Index korelace:**  $I = \pm\sqrt{I^2}; \quad I \in \langle -1; 1 \rangle$

## **2. Testové kritérium F**

- za vhodnější je považována ta funkce, u níž je hodnota F vyšší.

## **3. Reziduální součet čtverců a reziduální rozptyl**

**Reziduální součet čtverců:**  $S_R = \sum_{i=1}^n (y_i - Y_i)^2$

- za vhodnější považujeme funkci, která má reziduální součet čtverců nižší. Reziduální součet čtverců lze použít pouze tehdy, když srovnáváme funkce se stejným počtem parametrů.

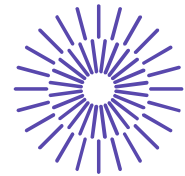
**Reziduální rozptyl:**  $s_R^2 = \frac{S_R}{n-p}$

- za vhodnější považujeme funkci, která má reziduální rozptyl nižší. Reziduální rozptyl můžeme použít vždy, bez ohledu na to, kolik parametrů mají srovnávané regresní funkce.

## **Korelační analýza**

- zabývá se především intenzitou vzájemného vztahu proměnných;
- je základní metodou měření síly lineární závislosti číselných proměnných;
- „correlatió“ = vzájemná souvislost (z lat.).

*! Z výpočetních a interpretačních hledisek se regresní a korelační analýza prolínají, nelze mezi nimi stanovit ostrou hranici.*



**Sdružené regresní přímky**

$Y = a_{yx} + b_{yx}x$  popisuje závislost  $y$  na  $x$

$X = a_{xy} + b_{xy}y$  popisuje závislost  $x$  na  $y$

$a_{yx} = \bar{y} - b_{yx}\bar{x}$        $b_{yx} = \frac{s_{xy}}{s_x^2}$

$a_{xy} = \bar{x} - b_{xy}\bar{y}$        $b_{xy} = \frac{s_{xy}}{s_y^2}$

1.  $b_{yx} = b_{xy} = 0 \Rightarrow$   $x$  a  $y$  jsou korelačně nezávislé; sdružené regresní přímky svírají pravý úhel.

2.  $b_{yx} = \frac{1}{b_{xy}} \Rightarrow$   $x$  a  $y$  jsou úplně závislé; sdružené regresní přímky svírají nulový úhel

(splývají).

**Míry těsnosti lineární závislosti**

**Koeficient determinace:**  $r_{yx}^2 = r_{xy}^2 = b_{yx} \cdot b_{xy} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_{xy}}{s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2}$  ;       $r_{xy}^2 \in \langle 0; 1 \rangle$

**Koeficient korelace:**  $r_{yx} = r_{xy} = \sqrt{r_{yx}^2} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(x^2 - \bar{x}^2) \cdot (y^2 - \bar{y}^2)}}$  ;       $r_{xy} \in \langle -1; 1 \rangle$

- měří sílu lineární závislosti, nikoli závislosti obecně (lineární závislost = korelovanost).

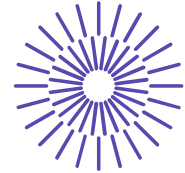
- koeficient je symetrický.

**Interpretace:**

1. znaménko +/- udává směr závislosti:

$r_{xy} > 0 \Rightarrow$  přímá závislost

$r_{xy} < 0 \Rightarrow$  nepřímá závislost



2.  $|r_{xy}|$  udává sílu závislosti:

$$r_{xy} = 0 \quad \Rightarrow \quad \text{lineární nezávislost}$$

$$|r_{xy}| = 1 \quad \Rightarrow \quad \text{funkční (úplná) závislost}$$

$$|r_{xy}| \rightarrow 0 \quad \Rightarrow \quad \text{slabá lineární závislost}$$

$$|r_{xy}| \rightarrow 1 \quad \Rightarrow \quad \text{silná lineární závislost}$$

### Test hypotézy o nulové hodnotě korelačního koeficientu

1)  $H_0 : \rho_{yx} = 0$  (lineární nezávislost  $x$  a  $y$ )

$$H_1 : \text{non } H_0$$

2) **Testové kritérium:**

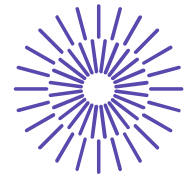
$$t = \frac{r_{yx} \cdot \sqrt{n-2}}{\sqrt{1-r_{yx}^2}}; \quad \text{Statistika } t \text{ má při platnosti } H_0 \text{ rozdělení } t \text{ s } (n-2) \text{ stupni volnosti.}$$

3) **Kritický obor:**

$$W \equiv \left\{ t; t \leq t_{\frac{\alpha}{2}}(n-2) \cup t \geq t_{1-\frac{\alpha}{2}}(n-2) \right\}$$

4) **Závěr testu:**

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme  $H_0$  a přijímáme  $H_1$ , tzn. prokázali jsme hypotézu o lineární závislosti proměnných  $x$  a  $y$ .



## Pořadová korelace

- Pokud chceme získat rychlou představu o síle závislosti mezi 2 kvantitativními znaky nebo určit závislost mezi pořadími znaků, nahradíme původní hodnoty  $x_i$  a  $y_i$  jejich pořadovými čísly  $i_x$  a  $i_y$  podle toho, která místa hodnoty zaujímají v uspořádané řadě.

**Spearmanův koeficient pořadové korelace:**  $r_s = 1 - \frac{6 \sum_{i=1}^n (i_x - i_y)^2}{n(n^2 - 1)}$  ;  $r_s \in \langle -1; 1 \rangle$

- varianta korelačního koeficientu;
- měří sílu lineární závislosti dvou pořadí.

**Interpretace:** stejná jako u korelačního koeficientu.

## Test hypotézy o nezávislosti pořadovou korelací

1)  $H_0 : \rho_s = 0$  (nezávislost pořadí)

$$H_1 : \text{non } H_0$$

2) **Testové kritérium:**

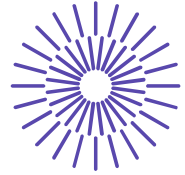
$$t = \frac{r_s \cdot \sqrt{n-2}}{\sqrt{1-r_s^2}} ; \quad \text{Statistika } t \text{ má při platnosti } H_0 \text{ rozdělení } t \text{ s } (n-2) \text{ stupni volnosti .}$$

3) **Kritický obor:**

$$W \equiv \left\{ t; t \leq t_{\frac{\alpha}{2}}(n-2) \cup t \geq t_{1-\frac{\alpha}{2}}(n-2) \right\}$$

4) **Závěr testu:**

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme  $H_0$  a přijímáme  $H_1$ , tzn. prokázali jsme hypotézu o lineární závislosti obou pořadí.



## **Vícenásobná lineární regrese**

- zkoumáme závislost proměnné  $y$  na dvou či více vysvětlujících proměnných (regresorech)

$x_1, x_2, \dots, x_k$ ;

- volba typu regresní funkce je obtížná; vhodné použití statistických programů;

- nejčastěji proto volíme lineární regresní funkci.

***Teoretická vícenásobná lineární regresní funkce:***

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_k x_k .$$