

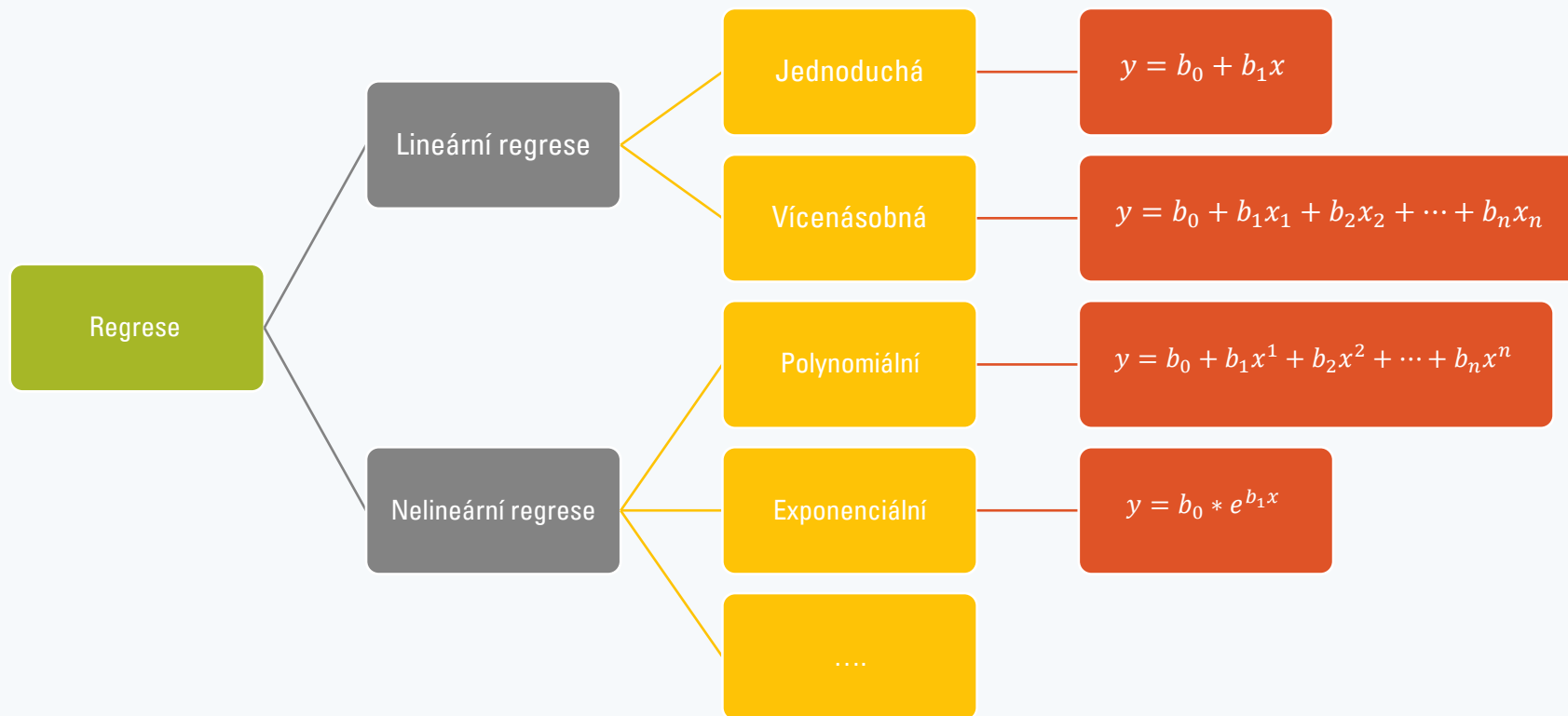


# Testy v regresi. Validace regrese

Pavla Pecherková

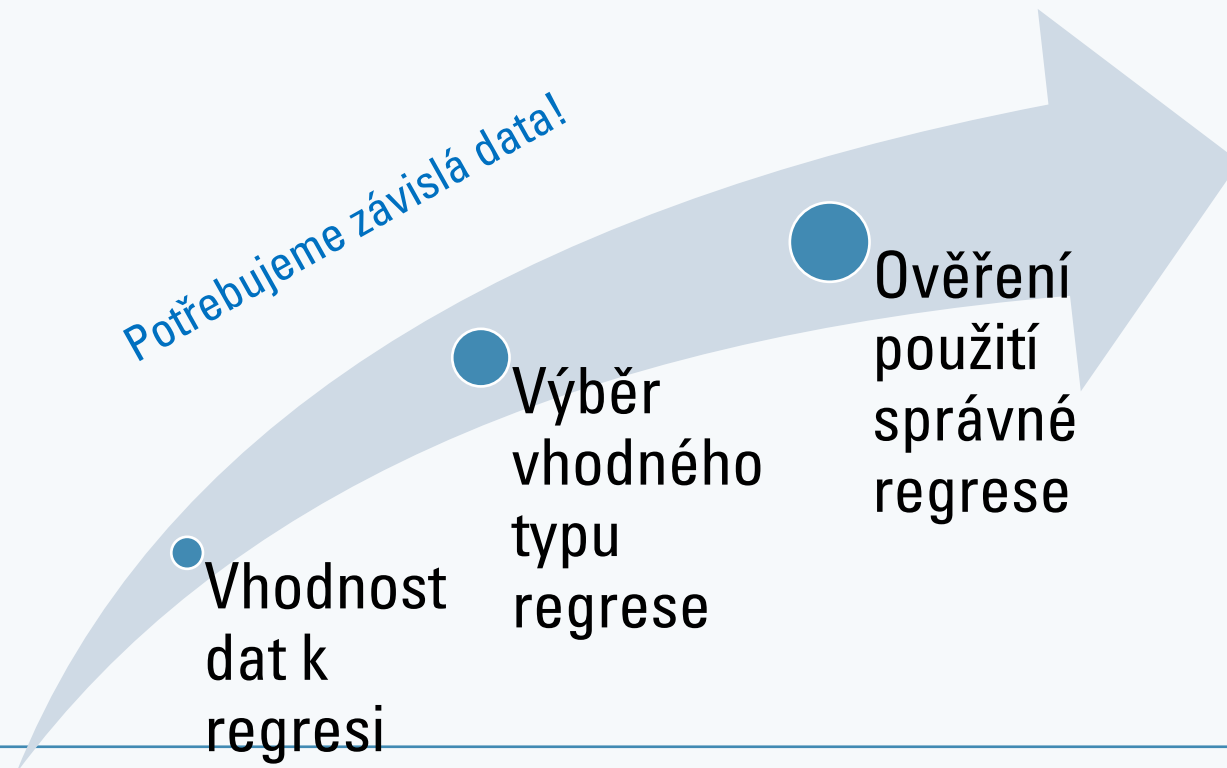
29. 4. 2021

# Druhy regresí



# Postup při zpracování regrese

- Jak vybrat jakou regresi použiji a zvolila jsem správně? Držíme se následujícího postupu:



# Vhodnost dat k regresi

Data jsou vhodná k regresi, pokud jsou závislá → korelační koeficient nesmí být roven 0



Potřebujeme zamítnout

## PEARSONŮV TEST KOR. KOEFICIENTU

Předpoklad:

1. Spojitá data,
2. předpokládáme  $N(\mu, \sigma^2)$ ,
3. párové výběry.

- $H_0$ : náhodné veličiny jsou lineárně nezávislé
- $H_A$ : náhodné jsou lineárně závislé

## Spearmanův test kor. koeficientu

Předpoklad:

1. Spojitá data,
2. párové výběry.

- $H_0$ : náhodné veličiny jsou nezávislé
- $H_A$ : náhodné jsou závislé

Lineární (jednoduchá)  
regrese

Vícenásobná lineární  
regrese

Nelineární  
polynomiální regrese

Nelineární  
exponenciální regrese

# Výběr vhodného typu regrese

Lineární (jednoduchá)  
regrese

Vícenásobná lineární  
regrese

Nelineární polynomiální  
regrese

Nelineární exponenciální  
regrese

# Regresní analýza

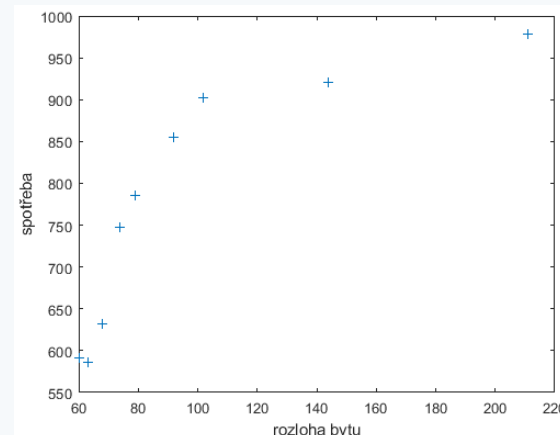
- Sledujeme měsíční spotřebu elektřiny v kWh vzhledem k rozloze domácnosti v m<sup>2</sup>.

| Velikost bytu | 60  | 63  | 68  | 74  | 79  | 92  | 102 | 144 | 211 |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| spotřeba      | 591 | 586 | 632 | 747 | 785 | 855 | 902 | 920 | 978 |

*Jsou zvolená data vhodná pro regresi? A pokud ano, jaká bude hodnota regresních koeficientů?*

1. Jsou data vhodná k regresi?

- Normalita dat – rozhodneme se, zda použijeme Pearsonův nebo Spearmanův test korelačního koeficientu
- Normalitu u prvního výběru (velikost bytu) zamítáme, proto použijeme Spearmanův test a nelineární regresi



```
[~,p_vb]=adtest(velikost_bytu)
```

```
p_vb = 0.0202
```

```
[~,p_s]=adtest(spotreba)
```

```
p_s = 0.4945
```

```
[r,p]=corr(velikost_bytu',spotreba','type','Spearman')
```

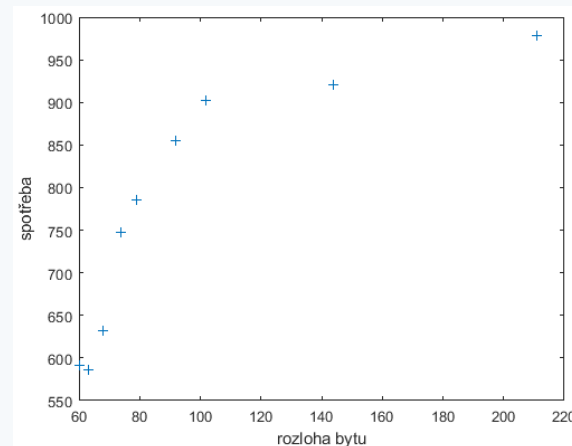
```
r = 0.9833
```

```
p = 4.9603e-05
```

Kterou nelineární regresi? Není zadání – zkusíme poznat z grafu

- Určování regrese:

1. Body nevypadají na exponenciální regresi → polynomiální regrese
2. Polynomiální regrese – 2. nebo 3. řádu? (sleduji počet „zlomů“)



mdl\_spotreba =  
Linear regression model:  
 $y \sim 1 + x1 + x1^2$

Estimated Coefficients:

|             | Estimate  | SE        | tStat   | pValue    |
|-------------|-----------|-----------|---------|-----------|
| (Intercept) | 31.821    | 137.72    | 0.23106 | 0.82494   |
| x1          | 11.705    | 2.4529    | 4.7719  | 0.0030883 |
| x1^2        | -0.034646 | 0.0090919 | -3.8106 | 0.0088576 |

Number of observations: 9, Error degrees of freedom: 6  
Root Mean Squared Error: 52.7  
R-squared: 0.905, Adjusted R-Squared: 0.873  
F-statistic vs. constant model: 28.6, p-value = 0.000859

2. řád

mdl\_spotreba =  
Linear regression model:  
 $y \sim 1 + x1 + x1^2 + x1^3$

Estimated Coefficients:

|             | Estimate   | SE         | tStat   | pValue     |
|-------------|------------|------------|---------|------------|
| (Intercept) | -1011.5    | 213.24     | -4.7436 | 0.0051338  |
| x1          | 40.78      | 5.799      | 7.0323  | 0.00089755 |
| x1^2        | -0.27905   | 0.048063   | -5.806  | 0.0021376  |
| x1^3        | 0.00061833 | 0.00012117 | 5.1028  | 0.0037612  |

Number of observations: 9, Error degrees of freedom: 5  
Root Mean Squared Error: 23.2  
R-squared: 0.985, Adjusted R-Squared: 0.976  
F-statistic vs. constant model: 107, p-value = 5.88e-05

3. řád

Porovnáme  
RMSE  
 $23.2 < 52.7$   
↓  
3. řád je lepší

Ale je vůbec polynomiální regrese vhodná? Dá se pro tato data použít? To nám řekne F-test

# Validace regrese

## F-test podílu vysvětleného a nevysvětleného rozptylu

- Poprvé jsme o něm mluvili u analýzy rozptylu (ANOVA).
- Budeme porovnávat odchylky mezi skutečnými (naměřenými)  $y$  a očekávanými (predikovanými)  $\hat{y} = y_p$ .
  - Vysvětlená odchylka = odchylka predikce od průměru ( $\hat{y}_i - \bar{y}$ )
    - Nemůže být stejná, protože to by predikovaná hodnota musela být vodorovná
  - Nevysvětlená odchylka = odchylka skutečné hodnoty od predikce ( $y_i - \hat{y}_i$ )
    - Nevíme, proč to není stejné – je to šum, nebo je to špatným typem regrese?
- Statistika:

$$T = \frac{(n-2) \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \sim F(n-1)$$

- $H_0$ : zvolená regrese je nevhodná
- $H_A$ : zvolená regrese je vhodná

⇒ Chceme zamítnout!

Předchozí příklad

```
mdl_spotreba =
Linear regression model:
y ~ 1 + x1 + x1^2 + x1^3
```

|             | Estimate   | SE         | tStat   | pValue     |
|-------------|------------|------------|---------|------------|
| (Intercept) | -1011.5    | 213.24     | -4.7436 | 0.0051338  |
| x1          | 40.78      | 5.799      | 7.0323  | 0.00089755 |
| x1^2        | -0.27905   | 0.048063   | -5.806  | 0.0021376  |
| x1^3        | 0.00061833 | 0.00012117 | 5.1028  | 0.0037612  |

Number of observations: 9, Error degrees of freedom: 5  
 Root Mean Squared Error: 23.2  
 R-squared: 0.985, Adjusted R-Squared: 0.976  
 F-statistic vs. constant model: 107, p-value = 5.88e-05

Zamítáme hypotézu, že zvolená regrese byla nevhodná.

# Validace regrese

## Test bělosti reziduí / Durbin-Watson test

Další z testů vhodnosti použité metody – jsou založeny na sledování reziduí (rozdílu mezi predikovanou hodnotou a skutečnou hodnotou), tedy  $e_i = y_i - \hat{y}_i$

### • Test bělosti reziduí

- $H_0$ : rezidua jsou nezávislá → regresní metoda je vhodná
- $H_A$ : rezidua nejsou nezávislá → regresní metoda **není vhodná**

### • Statistika:

$$T = \frac{2b - (n - 2)}{\sqrt{n - 1}} \sim N(0,1)$$

$$\text{kde } b = \sum_{i=1}^n b_i \quad \text{a} \quad b_i = e_i - \underbrace{e_{0,5}}_{\text{medián}}$$

### • Durbin-watson test

- $H_0$ : rezidua jsou nekorelovaná → regresní metoda je vhodná
- $H_A$ : rezidua jsou autokorelovaná → regresní metoda **není vhodná**

### • Statistika:

$$T = \frac{\sum_{i=1}^{n-1} (e_{i+1} - e_i)^2}{\sum_{i=1}^n e_i^2}$$

- kde  $n$  je počet pozorování.

Matlab Test na bělost reziduí neobsahuje v základu → jako alternativu k f-test používáme Durbin-Watson test.

**U obou testů nechceme zamítnout hypotézu (rozdíl proti f-test, kde chceme).**

# Vícenásobná lineární regrese

- U této regrese není možné otestovat korelaci, protože máme víc jak 2 výběry. Správně by měla být normalita dat, ale i to není zárukou pro správný výpočet. Na druhou stranu i „nenormalita“ dat nemusí vždy zaručovat, že data nejsou vhodná k regresi (pokud nezávisle proměnná  $x$  bude rok, tak to není normální rozdělení a přesto mohou být data vhodná k regresi).
- Zjednodušeně lze říci, že pokud mám víc nezávisle proměnných, není třeba testovat normalitu dat a uděláme rovnou vícenásobnou lineární regresi.

Měli bychom dodržet základní pravidla:

1. Nezávisle proměnné jsou **vzájemně nezávislé**.
2. Nezávisle proměnné jsou párové.
3. Máme spojitá data.

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

*Příklad:*

*Chceme zjistit, co ovlivňuje cenu čokolády. Máme 4 nezávisle proměnné: ( $x_1$ ) cukr, ( $x_2$ ) kakao, ( $x_3$ ) mléko a ( $x_4$ ) sušené mléko.*

*(nezávislost je mezi všemi veličinami mimo mléko a sušeného mléka - viz Matlab, ale pro názornost, tam necháme obě).*

```
mld_cokolada=fitlm([cukr' kakao' mleko' sus_mleko'],cokolada')
```

```
mld_cokolada =  
Linear regression model:  
y ~ 1 + x1 + x2 + x3 + x4
```

Estimated Coefficients:

|             | Estimate   | SE       | tStat    | pValue    |
|-------------|------------|----------|----------|-----------|
| (Intercept) | 10.587     | 4.2123   | 2.5135   | 0.030729  |
| x1          | -0.27451   | 0.1047   | -2.6218  | 0.025519  |
| x2          | 0.1343     | 0.039194 | 3.4266   | 0.0064754 |
| x3          | 0.66679    | 0.3092   | 2.1565   | 0.056443  |
| x4          | -0.0059329 | 0.03616  | -0.16407 | 0.87294   |

```
Number of observations: 15, Error degrees of freedom: 10  
Root Mean Squared Error: 1.3  
R-squared: 0.878, Adjusted R-Squared: 0.829  
F-statistic vs. constant model: 18, p-value = 0.000146
```

$H_0$ : parametr regrese  $b_i = 0$ , je možné, že parametry  $p > 0,05$  lze vynechat. Redukce má pravidla, ale to je nad rámec tohoto kurzu. Na první analýzu stačí informace o  $p$ .

$p > 0,05$

Vybraná metoda je vhodná.

# Postup při regresi

## 1. Test normality u obou výběrů

- a. zamítám  $H_0$  (data pochází z normálního rozdělení) → nelineární regrese
- b. nezamítám  $H_0$  (data pochází z normálního rozdělení) → lineární regrese

## 2. Test vhodnosti dat k regresi

- a. zamítám  $H_0$  (data jsou nezávislá) → pokračuji k bodu 3
- b. nezamítám  $H_0$  (data jsou nezávislá) → KONEC

## 3. Zvolení metody regresní analýzy na základě bodu 1 případně grafu

- a. Zjištění regresních koeficientů ( $b_0, b_1, \dots, b_n$ )
- b. U nelineární regrese – možnost porovnání RMSE

## 4. Predikce

- a. hledaná predikce v bodech,
- b. predikce v bodech závisle proměnných

## 5. Test vhodnosti použité metody regresní analýzy

- a. zamítám  $H_0$  (nebyla zvolena správná metoda) → KONEC (výsledky mohu použít)
- b. nezamítám  $H_0$  (nebyla zvolena správná metoda) → vrátím se ke kroku 3 a zkusím jinou metodu

- Výjimka: Vícenásobná lineární regrese – začíná u 3