Přednáška 10 – Testy v regresi. Validace regrese

Minulý týden v přednášce 9 jsme se podívali na testy nezávislosti, kde jsme si poznamenali, že některé z testů slouží k určování, zda naměřená data jsou vhodná k regresní analýze. Připomeňme si, že to byly testy nezávislosti určené pro spojité veličiny, které jsme dělili na parametrické a neparametrické podle toho, zda výběry splňují předpoklad normality či nikoliv.

Nulová hypotéza testu nezávislosti obecně tvrdí, že veličiny jsou nezávislé. Tedy zamítnutí nulové hypotézy logicky znamená, že mezi veličinami je vazba, tím pádem jsou data vhodná k regresi.

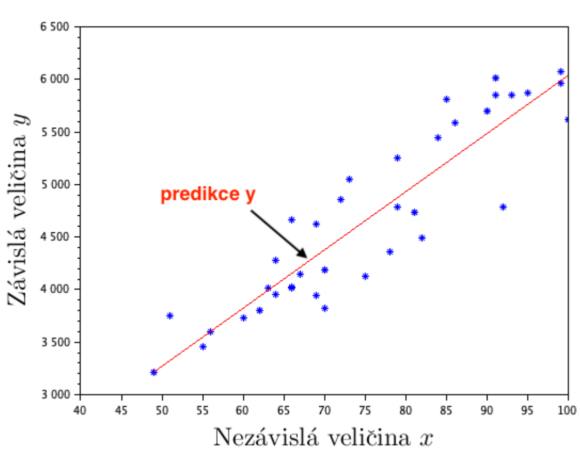
V případě zamítnutí parametrického Pearsonova testu považujeme data za vhodná k lineární regresi. Pokud zamítneme jeho neparametrickou alternativu Spearmanův test, znamená to, že data jsou vhodná k nelineární regresi – polynomiální, exponenciální, atd (viz přednáška 4). V následující tabulce jsou testy na vhodnost k regresní analýze, které budeme používat (je také dostupná na webu na odkazu Jak zvolit test hypotéz).

Testy na vhodnost k regresní analýze	
Parametrické	Neparametrické
Pearsonův test – **pearson_test** Předpoklady: $N(\mu, \sigma^2)$, párové výběry H_0: jsou nezávislé Pokud zamítáme: data jsou vhodná k lineární regresi $y = b_0 + b_1 x$	Spearmanův test – **spearman_test** Předpoklady: bez $N(\mu, \sigma^2)$, párové výběry H_0: jsou nezávislé Pokud zamítáme: data jsou vhodná k nelineární regresi $y = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n$ $y = b_0 \exp\{b_1 x\}$

Validace regrese

Pokud máme data vhodná k regresi, znamená to, že můžeme použít data pro regresní analýzu. Neznamená to ale, že takový pokus bude zaručeně úspěšný. Proto po provedení regrese je vhodné výsledky ověřit a otestovat, jestli vybraná regresní metoda vyhovovala naměřeným datům. Proces ověření výsledků regrese se nazývá validace regrese.

Základem validace regrese je porovnání hodnot naměřené závislé veličiny y a její predikce, tj., hodnot, které leží na regresní přímce (v případě lineární regrese) nebo křivce (v případě nelineární regrese). Připomeňme si, jak vypadá lineární regrese na obrázku:



Na obrázku vidíme data nezávislé veličiny x a závislé veličiny y vykreslená proti sobě. Červená přímka je lineární regrese, která obsahuje hodnoty predikce $\hat{y}$ po dosažení odhadů regresních koeficientů do regrese:

$$\underbrace{\hat{y}_i}_{\text{hodnoty na přímce}} = \overbrace{\hat{b}_0 + \hat{b}_1}^{\text{odhad}} \underbrace{x_i}_{\text{data}}.$$

V průběhu validace porovnáváme hodnoty y a $\hat{y}$, tj., testujeme, zda predikce ukazuje správně trend vývoje dat. Jsou na to speciální testy hypotéz – my probereme ty nejčastěji používané.

Testy hypotéz pro validaci regrese

V následující tabulce jsou uvedeny testy validace regrese, které budeme používat (tabulka je také dostupná na webu na odkazu Jak zvolit test hypotéz). Neexistuje striktní pokyn, jaký z těchto testů zvolit, tj., pro validaci se hodí oba. Obecně se více doporučuje F-test podílu vysvětleného a nevysvětleného rozptylu, který se nachází v levém sloupci – tento test je silnější.

F-test podílu vysvětleného a nevysvětleného rozptylu	Test nezávislosti reziduí
f_test_pred	**wz_test**
H_0 : zvolená regrese je nevhodná	H_0 : zvolená regrese je vhodná
Pokud zamítáme: regrese byla vhodná	Pokud zamítáme: regrese nebyla vhodná

Probereme každý z testů podrobněji i s příkladem.

F-test podílu vysvětleného a nevysvětleného rozptylu

F-test podílu vysvětleného a nevysvětleného rozptylu (**f_test_pred**) je dost obecný test, používali jsme ho pro analýzu rozptylu u testu Anova. Tady ho využijeme pro testování shody naměřených hodnot závislé veličiny y a její predikce $\hat{y}$. Výpočet statistiky testu je založen na využití následujícího vztahu y a $\hat{y}$:

$$\underbrace{\underbrace{y_i}_{\text{data}} - \underbrace{\bar{y}}_{\text{průměr}}}_{\text{odchylka dat od průměru}} = \underbrace{\underbrace{\underbrace{y_i - \hat{y}_i}_{\text{odchylka dat od predikce}}}_{\text{reziduum } e_i = y_i - \hat{y}_i}}_{\text{nevysvětlená odchylka}} + \underbrace{\underbrace{\hat{y}_i - \bar{y},}_{\text{odchylka predikce od průměru}}}_{\text{vysvětlená odchylka}}$$

kde jsme k odchylce naměřených dat od průměru pouze přidali a odečetli predikci $\hat{y}_i$ v každém bodu. Dále odchylka dat od predikce (reziduum) tvoří nevysvětlenou odchylku – neumíme ji vysvětlit, protože hodnoty predikce y_i by měly být co nejblíže naměřeným datům y. Odchylku predikce od průměru umíme vysvětlit – $\bar{y}$ je číslo a hodnoty predikce $\hat{y}_i$ leží na přímce, která není vodorovná. Proto je to vysvětlená odchylka.

Test používá statistiku

$$F = \frac{(n-2)\text{vysvětlená odchylka}}{\text{nevysvětlená odchylka}} = \frac{(n-2)\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \sim \text{ Fisherovo rozdělení.}$$

Test je pouze pravostranný. Při použití F-testu je potřeba dát velký pozor na nulovou hypotézu, která zní:

H_0 : zvolená regrese je **nevhodná**,

H_A : je vhodná.

Proto je pro nás výhodněji nulovou hypotézu zamítnout. Pokud bychom ji nezamítli, znamenalo by to, že musíme použít jinou regresní metodu.

Příklad: Sledujeme měsíční spotřebu elektřiny a rozlohu několika domácností. Data jsou v tabulce, kde v prvním řádku je velikost bytu v m^2, v druhém řádku je spotřeba elektřiny v kWh. Zajímá nás, zda tato data jsou vhodná k regresi. Pokud ano, použijeme je a následně ověříme, zda zvolený typ regrese byl vhodný.

m^2	60	63	68	74	79	92	102	144	211	60	63	68	74	79	92	102	144	211
kWh	591	586	632	747	785	855	902	920	978	591	586	632	747	785	855	902	920	978

Řešení: Tady by se dalo uvažovat o lineární regresi – zdálo by se, že čím větší byt je, tím vyšší je spotřeba. Jenomže ve velkém bytě mohou bydlet nějací šetřílci, kteří mají všude LED světla a doma se objevují pouze večer. Naopak menší byt může patřit rodině s dvěma malými detmi, kde každou chvíli pere pračka, hodně se vaří, atd., tj., spotřeba je mnohem vyšší.

Nejprve otestujeme data na normalitu. Pro první z výběrů jsme zamítli předpoklad normality, proto použijeme neparametrický Spearmanův test (**spearman_test**), abychom otestovali, zda jsou data vhodná k regresi. Nulová hypotéza Spearmanova testu zní:

$$H_0: \quad \text{velikost bytu a spotřeba elektřiny jsou nezávislé,}$$
$$H_A: \quad \text{nejsou nezávislé.}$$

P-hodnota$= 2.841D-13$, je menší než hladina významnosti 0.05, proto zamítáme nulovou hypotézu, že data jsou nezávislá. To znamená, že mezi velikostí bytu a spotřebou elektřiny je vazba a data jsou vhodná k regresní analýze.

Jelikož jsme použili neparametrický Spearmanův test, data jsou vhodná k nelineární regresi. Zkusíme použít například polynomiální regresi 3.řádu:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3,$$

kde nezávislá veličina x je velikost bytu a závislá veličina y je spotřeba elektřiny. Odhadneme regresní koeficienty podle metody nejmenších čtverců (viz přednáška 4) pomocí funkce **pol_reg**:

$$y = -1011.5107 + 40.780131x - 0.2790548x^2 + 0.0006183x^3.$$

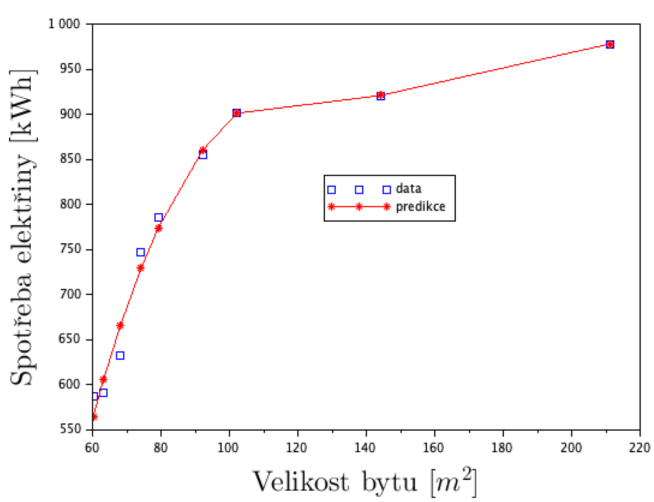
Teď spočítáme predikci spotřeby, tj., hodnoty na regresní křivce polynomiální regrese pomocí funkce **pol_pred**, kam dosadíme odhady regresních koeficientů a všechna data x. Dostaneme vektor predikce s hodnotami na křivce:

```
564.26004 604.68164 665.61287 728.67879 773.40157 859.8311 900.95798 920.68492 977.8911 564.26004
604.68164 665.61287 728.67879 773.40157 859.8311 900.95798 920.68492 977.8911
```

Je vidět, že hodnoty predikce odpovídají naměřeným hodnotám spotřeby v tabulce, ale nedokážeme říct, nakolik dobře. Proto využijeme test pro validaci regrese F-test podílu vysvětleného a nevysvětleného rozptylu (**f_test_pred**), tj., ověříme, zda zvolená regrese popisuje dobře vývoj dat. Řekneme si nulovou hypotézu:

$$H_0: \quad \text{zvolená regrese je } \textbf{nevhodná},$$
$$H_A: \quad \text{je vhodná.}$$

P-hodnota$= 6.156D-13 < 0.05$, takže zamítáme nulovou hypotézu, že zvolená regrese nebyla vhodná. To znamená, že jsme použili vyhovující – správnou metodu. Můžeme se o tom přesvědčit i na obrázku:



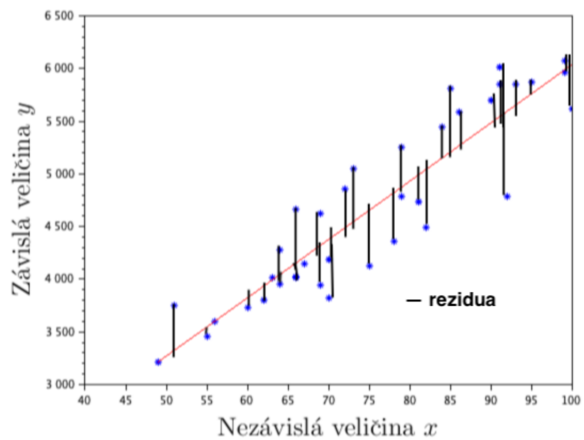
Test nezávislosti reziduí

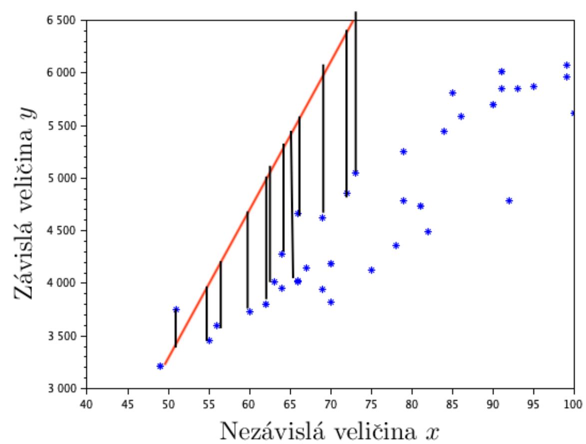
Test nezávislosti reziduí (občas test bělosti reziduí) (**wz_test**) použijeme pro validaci výsledků regrese. Připomeňme si, že rezidua

$$e_i = y_i - \hat{y}_i$$

jsou odchylky od regresní přímky v každém bodě (viz přednáška 4).

Test zkoumá, zda rezidua jsou nekorelovaná. Je založen na principu, že pokud byla regrese zvolena dobře, tj., správně ukazuje trend vývoje naměřených dat, rezidua by měla být kladná a záporná a pořadově nezávislá, jak můžeme například vidět na obrázku vlevo. Při správně zvolené regresní metodě by nemělo docházet k pouze kladným nebo pouze záporným hodnotám reziduí, jako například na obrázku vpravo. Tady je vidět, že regresní přímka neprochází daty, v důsledku čehož rezidua narůstají a jsou potom pouze záporná.





Statistika testu používá výpočet

$$b_i = e_i - \underbrace{\tilde{e}_{0.5}}_{\text{medián}}, \qquad b = \sum_{i=1}^{n} b_i.$$

Statistika je

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

Nulová hypotéza zní:

H_0 : rezidua jsou nezávislá, tj., zvolená regrese je **vhodná**,

H_A : není vhodná.

Všimněme si, že kvůli tomu, že test je založen na nekorelovanosti reziduí, je nulová hypotéza obrácená v porovnání s F-testem. Pro nás je výhodné ji nezamítnout. Pokud bychom ji zamítli, museli bychom zvolit jiný typ regrese.

Poznámka: Tomuto testu se také říká test bělosti reziduí, protože nezávislost jednotlivých prvků je jedna z vlastností bílého šumu. Test se používá i za jiným účelem, než pro validaci regrese.

Příklad: Sledujeme vývoj ceny kakaa a mléčné čokolády ročně v období 2004-2018. V tabulce jsou uvedeny ceny za 100g v Kč. Zajímá nás, zda jsou data vhodná k regresi, a pokud ano, potřebujeme ověřit výsledky regrese.

```
kakao=[41.605 39.289 35.377 38.481 43.206 53.298 56.667 58.240 47.463 43.460 ...
59.526 66.447 76.241 48.274 50.516];
coko=[19.65 19.75 17.78 19.42 20.58 21.91 22.17 23.58 22.22 21.32 25.63 ...
25.57 27.31 27.12 27.20];
```

Řešení: Nejdříve otestujeme data na normalitu, abychom zjistili, který z testů na vhodnost k regresi můžeme použít. Oba výběry pochází z normálního rozdělení, což znamená, že použijeme Pearsonův test (**pearson_test**). Nulová hypotéza Pearsonova testu zní:

H_0 : ceny kakaa a mléčné čokolády jsou lineárně nezávislé,
H_A : nejsou lineárně nezávislé.

P-hodnota$= 0.0008741 < 0.05$, proto zamítáme nulovou hypotézu, že data jsou lineárně nezávislá. To znamená, že data jsou vhodná k lineární regresi:

$$y = b_0 + b_1 x,$$

kde nezávislá veličina x je cena kakaa a závislá veličina y je cena mléčné čokolády. Pro výpočet regresních koeficientů lineární regrese použijeme funkci **lin_reg**:

$$y = 11.985553 + 0.2129387x.$$

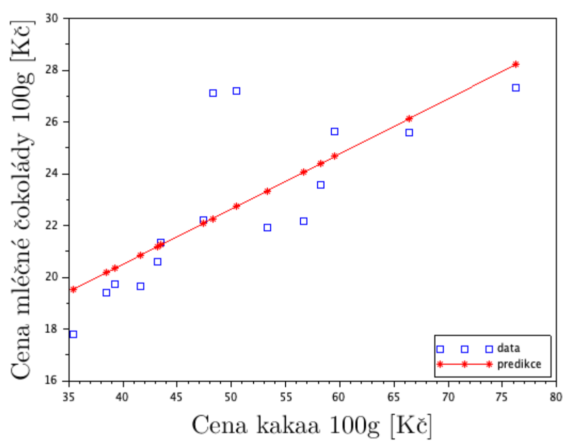
Pro výpočet hodnot regresní přímky, tj., predikce ceny mléčné čokolády, dosadíme odhady a data x do funkce **lin_pred**. Výsledná predikce je

```
20.844868 20.351702 19.518686 20.179647 21.185783 23.33476 24.052151 24.387103 22.092263 21.239869
24.660943 26.134691 28.220213 22.264956 22.742365
```

Abychom posoudili, zda je regresní přímka vhodná pro naměřená data, použijeme pro validaci regrese test nezávislosti reziduí (**wz_test**). Řekneme si nulovou hypotézu:

H_0 : rezidua jsou nezávislá, tj., zvolená regrese je **vhodná**,
H_A : není vhodná.

P-hodnota$= 0.2905273 > 0.05$, proto nezamítáme nulovou hypotézu, že zvolená regrese je vhodná, což vidíme na obrázku:



Poznámka: Jak je vidět, F-test a test nezávislosti reziduí mají prohozené nulové hypotézy. Můžeme použít jakýkoliv z testů, důležité je však znát nulovou hypotézu.