Marie Turčičová

Lineární regrese I

Obecně regresní model slouží k vysvětlení kolísání (variability) závisle proměnné Y pomocí chování nezávisle proměnné x (popř. více proměnných). Pokud se nám podaří takový vhodný model najít, můžeme dále:

- testovat závislost Y na x
- předpovídat střední hodnotu Y při zvoleném x

Asi nejjednodušší je uvažovat lineární regresní závislost, kdy Y je lineární funkcí x. Předpokládáme tedy, že střední hodnota Y závisí na x skrze model:

$$\mathrm{E}\,Y = \beta_0 + \beta_1 x \tag{1}$$

Než budeme pokračovat dále, ujasněme si terminologii, která se pro jednotlivé složky takového modelu používá (v závorce je vždy kurzívou uveden odpovídající anglický termín).

➔ Y je náhodná veličina a používají se pro ni označení: závisle proměnná (*dependent variable*), odezva (*response*), vysvětlovaná proměnná (*explained variable*)

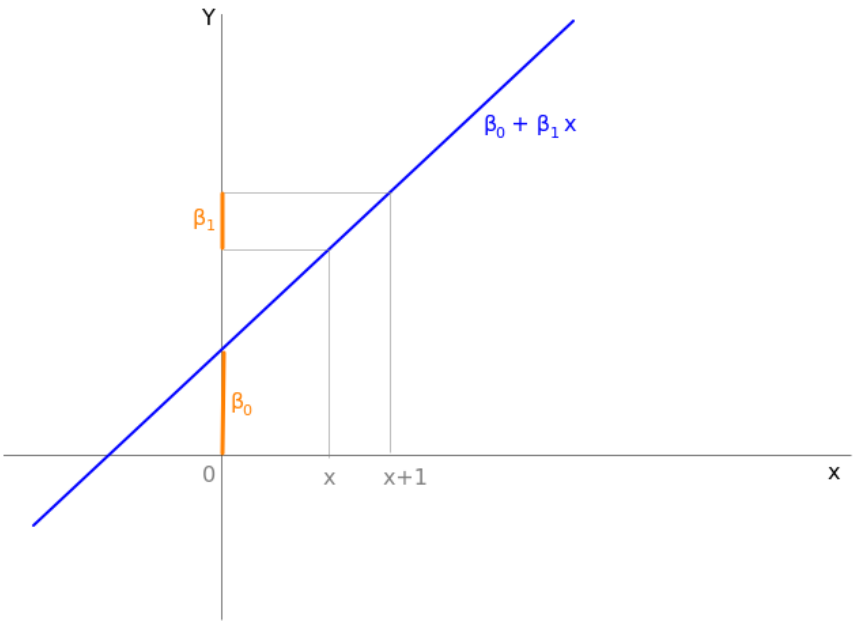
➔ x se považuje za přesně známé číslo (není to náhodná veličina) a nazývá se: nezávisle proměnná (*independent variable*), regresor (*regressor*), vysvětlující proměnná (*explanatory variable*)

➔ β_0 se nazývá absolutní člen (*intercept*) nebo také posunutí

➔ β_1 se nazývá směrnice nebo též sklon (*slope*)

1. Význam regresních koeficientů

Koeficienty β_0 a β_1 se souhrnně nazývají **regresní koeficienty** (*regression coefficients*) a udávají absolutní člen a směrnici příslušné regresní přímky (viz obrázek). Odtud je tedy evidentní i jejich význam. Parametr β_0 nám říká, jaká je střední hodnota Y pro $x = 0$. Parametr β_1 udává změnu střední hodnoty Y při jednotkové změně x, tj. s každou další jednotkou x vzrose $\mathrm{E}\,Y$ o β_1.



2. Model pro konkrétní data

Předpokládejme nyní, že máme k dispozici n realizací (Y, x), tj. máme data

$$\begin{aligned} Y_1, Y_2, \ldots, Y_n &= \text{nezávislá pozorování} \\ x_1, x_2, \ldots, x_n &= \text{známá čísla} \end{aligned}$$

Model (1) lze ekvivalentně přepsat do tvaru

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n \tag{2}$$

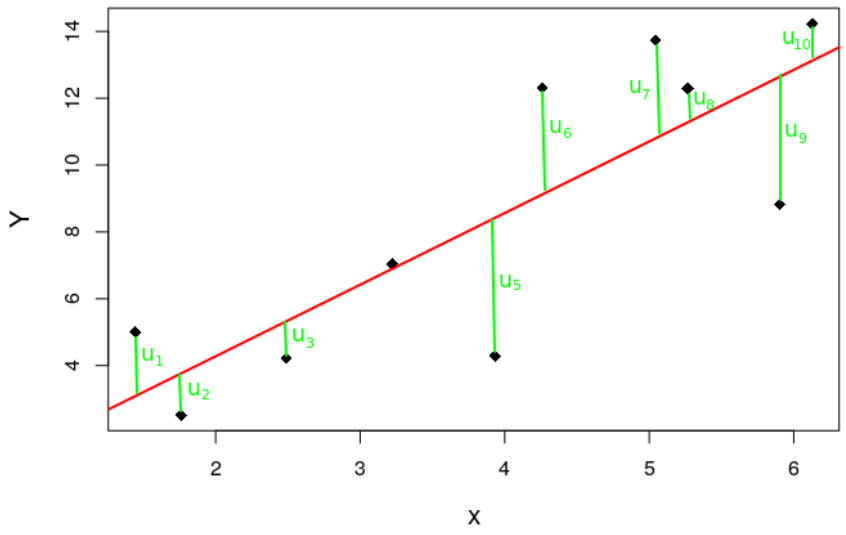
kde e_i je náhodná odchylka Y_i od střední hodnoty $\mathrm{E}\,Y$. Předpokládáme, že odchylky/chyby $e_1, e_2, \ldots, e_n$ mají stejný rozptyl σ^2, který není znám. Abychom mohli později testovat hypotézy o β_0 a β_1, budeme ještě potřebovat předpoklad, že $e_i \sim N(0, \sigma^2)$ pro všechna $i = 1, \ldots, n$. Lineární regresní model (2) má tedy celkem 3 parametry: β_0, β_1 a σ^2, které budeme muset odhadnout.

Při vysvětlování variability Y_i tvoří $\beta_0 + \beta_1 x_i$ systematickou složku, která je vysvětlena předpokládanou lineární závislostí, kdežto e_i je náhodná nevysvětlená složka.

3. Odhadování parametrů modelu

3.1 Odhad regresních koeficientů

Otázkou teď je - mám-li k dispozici konkrétní data, jak tam tu regresní přímku proložit? Dělá se to metodou nejmenších čtverců (*least squares method*), tj. přímka se proloží tak, aby součet zelených vzdáleností (umocněných na druhou) byl co nejmenší. Matematicky to lze zapsat tak, že β_0, β_1 odhadu-



jeme z podmínky, že

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{3}$$

má být minimální. Výsledné odhady parametrů β_0, β_1 si označíme b_0 a b_1. Lze spočítat[1], že

$$b_0 = \bar{Y} - b_1 \bar{x} \tag{4}$$

$$b_1 = \frac{\sum_{i=1}^{n} Y_i x_i - n\bar{Y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \tag{5}$$

[1] Možná si pamatujete ze střední školy, že minimum funkce se hledalo tak, že se příslušná funkce zderivovala a hledal se bod, kde je tato derivace rovná nule. Stejně to funguje i zde. Kdybyste výraz (3) zderivvoali podle β_0 a výsledek položili rovný nule, dostali byste přesně naše b_0. Stejně tak pro β_1 a b_1

kde $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ a $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Poznamenejme, že odhady b_0, b_1 se nazývají odhady metodou nejmenších čtverců, v anglické literatuře jako *least-squares estimators (LSE)*. Odhadnutá regresní přímka má rovnici $y = b_0 + b_1 x$ a odhadem regresního modelu (2) je:

$$Y_i = b_0 + b_1 x_i + u_i, \quad i = 1, \ldots, n. \tag{6}$$

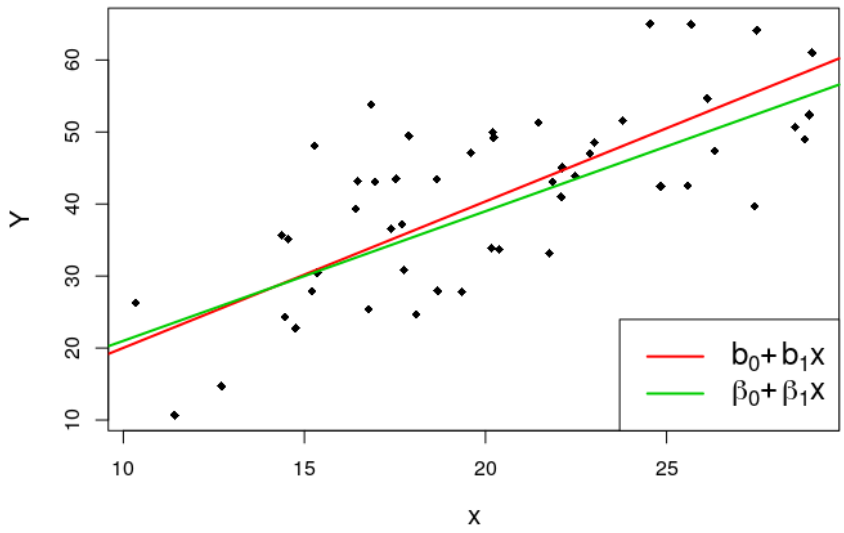
Body na regresní přímce $\hat{Y}_i = b_0 + b_1 x_i$ se nazývají odhadnuté, vyrovnané, nebo též **vyhlazené hodnoty**. Odhadem odchylek e_i z (2) jsou hodnoty $u_i = Y_i - \hat{Y}_i$, které se nazávají **rezidua**.

Hodnota dosaženého minima výrazu (3) se nazývá **reziduální součet čtverců** (jde opravdu a součet čtverců/kvadrátů reziduí)

$$S_e = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1} U_i^2 \tag{7}$$

a představuje nevysvětlenou variabilitu Y.

Poznámka 1 *Uvědomte si, prosím, že skutečná regresní přímka $\beta_0 + \beta_1 x$ a odhadnutá regresní přímka $b_0 + b_1 x$ jsou obecně dvě různé přímky, které se mohou (někdy i výrazně) lišit (viz obrázek níže). Skutečná regresní přímka nám zůstává neznámá.*



3.2 Odhad σ^2

Odhadem parametru σ^2 je tzv. **reziduální rozptyl**

$$s^2 = \frac{S_e}{n-2}. \tag{8}$$

4. Prokazování závislosti

Až v této sekci využijeme předpoklad, že náhodné chyby $e_1, e_2, \ldots, e_n$ mají normální rozdělení $N(0, \sigma^2)$.

Pomocí sestaveného modelu lze kromě jiného testovat závislost odezvy Y na regresoru x. V našem modelu (1),(2) nezávislost Y na x odpovídá tomu, že $\beta_1 = 0$. Je-li totiž $\beta_1 = 0$, pak vliv x vymizí. Test nezávislosti Y na x je tedy vlastně testem hypotézy

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

Zvolenou hladinu testu si tradičně označíme α. Nikoho asi nepřekvapí, že příslušný test je založený na odhadu β_1, tj. na b_1. Je-li b_1 velké, můžeme H_0 zamítnout. Konkrétně se test provádí pomocí statistiky

$$T = \frac{b_1}{\text{s.e.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}^2)}, \tag{9}$$

kde „s.e." značí směrodanout chybu (*standard error*) a s značí odmocninu z reziduálního rozptylu (8), tj. $s = \sqrt{\frac{S_e}{n-2}}$. Statistika T má v případě, že H_0 platí, t-rozdělení s $n-2$ stupni volnosti, tj. $T \sim t_{n-2}$ (a právě k důkazu tohoto tvrzení je potřeba normalita náhodných chyb $e_1, e_2, \ldots, e_n$).

Hypotézu H_0 zamítneme, pokud $|T| \geq qt_{n-2}\left(1 - \frac{\alpha}{2}\right)$.

Pokud H_0 zamítneme, můžeme říct, že „na hladině α je závislost Y na regresoru x průkazná".

5. Hodnocení kvality modelu

Sestavíme-li nějaký regresní model, je vždy potřeba nějak ohodnotit jeho kvalitu, tj. do jaké míry se nám pomocí něj podařilo kolísání Y vysvětlit. Než dojdeme ke vhodné charakteristice, kterou lze kvalitu modelu hodnotit, musíme provést následující úvahy.

Kdybychom ignorovali možnou závislost Y na x (tj. zvolili bychom $\beta_1 = 0$), nejlepším odhadem β_0 by byla průměrná hodnota závisle proměnné, tj. $\bar{Y}$. Reziduální součet čtverců takového modelu by byl

$$S_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{10}$$

a nazývá se **celkový součet čtverců**. Pokud připustíme závislost na x, tj. uvažujeme náš model (2), tak máme náš reziduální součet čtverců (7). Lze přitom ukázat, že

$$S_T - S_e = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 =: S_R \tag{11}$$

což se nazývá **regresní součet čtverců** a označili jsme si ho jako S_R. (Připomeňme, že $\hat{Y}_i = b_0 + b_1 x_i$ jsou vyhlazené hodnoty.)

Co se týče interpretace, tak celkový součet čtverců vyjadřuje celkovou variabilitu hodnot závisle proměnné Y. Regresní součet čtverců zase vyjadřuje variabilitu vyhlazených hodnot $\hat{Y}_i$, tedy tu část variability závisle proměnné, kterou se nám podařilo vysvětlit lineární závislostí. Na těchto kvantitách je vhodné založit hodnocení kvality modelu, protože dobrý model by měl vysvětlovat vysokou část celkové variability, tj. S_R/S_T by mělo být blízko 1. Přesně na tomto pozorování je založen **koeficient determinace**, který se obvykle značí R^2 a vypočte se jako

$$R^2 = \frac{S_R}{S_T} = \frac{\text{variabilita vysvětlená}}{\text{variabilita celková}} = 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita celková}} = 1 - \frac{S_e}{S_T} \tag{12}$$

Vzoreček $R^2 = 1 - \frac{S_e}{S_T}$ nejčastěji najdete v učebnicích. Dále existuje ještě korigovaný koeficient determinace, který bere v úvahu i počet pozorování

$$\bar{R}^2 = 1 - \frac{n-1}{n-2}\frac{S_e}{S_T}. \tag{13}$$

Oba tyto koeficienty jsou bezrozměrná čísla (tj. nemají žádné jednotky) a čím jsou blíže k jedné, tím je regresní závislost těsnější. Udávají, jakou část variability závisle proměnné Y jsme uvažovanou závislosti vysvětlili. Ekvivalentně můžeme říci:

- ➔ Závislostí na x jsme vysvětlili $R^2 \cdot 100\%$ variability Y.
- ➔ Kolísání Y lze z $R^2 \cdot 100\%$ vysvětlit závislostí na x.

Poznámka 2 *(pro zájemce) U lineární regrese je koeficient determinace R^2 roven druhé mocnině Pearsonova korelačního koeficientu $r_{y,x}^2$.*

6. Souvislost s analýzou rozptylu

Rozklad variability Y tvaru

$$S_T = S_R + S_e \tag{14}$$

by vám ale měl něco připomínat! Na tom byla přece založena analýza rozptylu!

V modelu analýzy rozptylu byla x kategoriální náhodná veličina a pro každou její hodnotu (kategorii) jsme uvažovali nějakou, obecně různou, střední hodnotu Y. V regresi je veličina x kvantitativní a zmíněné střední hodnoty nejsou libovolné, ale jsou vázány tím lineárním modelem. Regrese tedy velmi připomíná ANOVu, kde bychom si představili „spojité kategorie“ dané pomocí x_i. Na základě této představy je evidentní, že

- regresní součet čtverců $S_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ nahrazuje meziskupinovou variabilitu z ANOVy
- reziduální součet čtverců $S_e = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ nahrazuje vnitroskupinovou variabilitu z ANOVy

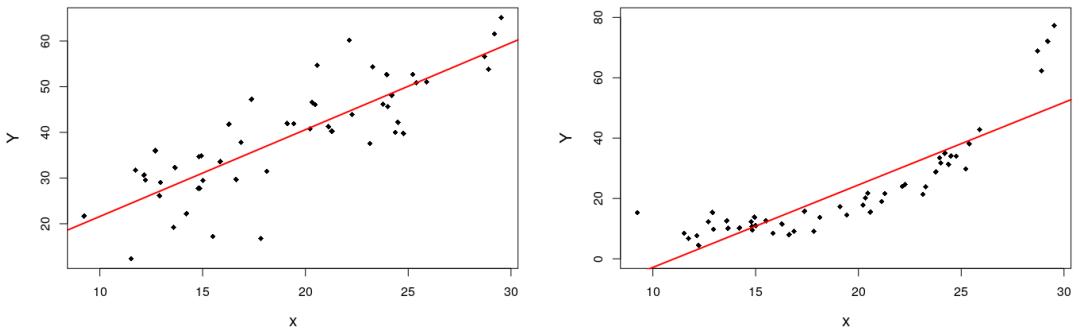
Jelikož ANOVA i lineární regrese jsou obojí testem nezávislosti Y na x (pouze jednou je x kategoriální a podruhé spojitá veličina), nikoho už asi nepřekvapí, že p-hodnota testu hypotézy $H_0 : \beta_1 = 0$ (proti $H_1 : \beta_1 \neq 0$) vyjde stejně jako p-hodnota v tabulce příslušné analýzy rozptylu. (To uvidíme později na konkrétním příkladu v R).

7. Ověřování předpokladů

Proložit si daty přímku metodou nejmenších čtverců není nic proti ničemu a nikdo vám v tom nemůže zabránit :o). Jakmile ale máte od svého regresního modelu nějaká vyšší očekávání, např. chcete testovat závislost Y na daných regresorech, musíte mít splněny veškeré předpoklady této metody. Pojďme si je nyní na závěr shrnout. Stejně jako u analýzy rozptylu, některé z nich bude možné ověřit až po odhadnutí modelu.

1. **Nezávislost** veličin $Y_1, Y_2, \ldots, Y_n$.
   To je předpoklad který musí být zajištěn vhodnou organizací pokusu a dodatečně už se s ním nedá nic dělat. Problém s nezávislostí může nastat například tam, kde působí čas (tj. hodnoty $Y_1, \ldots, Y_n$ tvoří časovou řadu a podobně).

2. Zvolený **tvar modelu** je správný.
   Tento předpoklad se ověřuje na základě různých diagnostických grafů (uvidíme v R), kdy zkoumáme, jestli je tvar uvažované závislosti vhodný. Nemělo by se například stát, že body tvoří kolem regresní přímky podkovu a podobně.

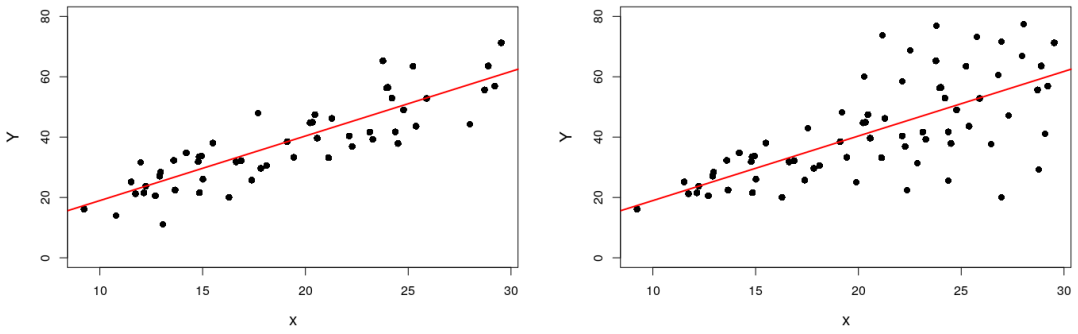


(a) Ano (b) Ne

Obrázek 1: Vhodný tvar modelu.

3. Náhodné chyby $e_1, \ldots, e_n$ mají **normální rozdělení** $N(0, \sigma^2)$.
   Odhadem náhodných chyb $e_1, \ldots, e_n$ jsou rezidua $u_1, \ldots, u_n$, proto se tento předpoklad testuje pomocí těchto reziduí, nebo jejich modifikace (standardizovaná rezidua), na která se aplikuje Shapiro-Wilkův test.

4. Náhodné chyby $e_1, \ldots, e_n$ mají shodný rozptyl σ^2 - tomu se říká **homoskedasticita**[2].
   Toto se opět ověřuje z diagnostických grafů. Data by kolem regresní přímky měla tvořit homogenní pás, nikoli například trychtýř... (viz Obrázek 2). Rigorózně lze homoskedasticitu ověřit pomocí Breuschova-Paganova testu.



(a) Ano (b) Ne

Obrázek 2: Homoskedasticita.

Poznámka 3 *Řadu problému s nesplněním předpokladů lze odstranit vhodnou transformací závisle či nezávisle proměnné (hodí se např. logaritmus, druhá mocnina apod.).*

[2]Naopak nekonstantnost rozptylu se nazývá heteroskedasticita.