Lineární regresní model

Mgr. Zdeňka Geršlová

Jednoduchá lineární regrese

- zkoumáme **lineární** vztah dvou proměnných
- **Y ... závislá** (vysvětlovaná) proměnná, odezva (outcome, response, dependent variable)
  X ... nezávislá (vysvětlující) proměnná, prediktor, regresor (predictor, explanatory or independent variable)
- Proč? Chceme získat předpis, pomocí kterého budeme schopni **předpovědět hodnotu jedné proměnné ze znalosti hodnoty jiné proměnné**, pokud mezi těmito dvěma proměnnými existuje příčinná souvislost

Matematický zápis

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

kde β_0, β_1 jsou regresní koeficienty, které odhadujeme z dat

- β_0 … konstantní člen (intercept, posun přímky ve směru osy y)
  … střední hodnota predikce při nulových hodnotách všech ostatních regresorů
- β_1 … směrnice přímky (sklon přímky, gradient)
  … rozdíl v predikci, jestliže vzroste příslušný regresor o jednu jednotku

ε_i … chybová (reziduální) složka ("šum")

Kvalita modelu

Cíl: proložit (naměřenými) daty přímku, která co nejlépe vystihuje vztah mezi proměnnými - používáme metodu nejmenších čtverců (chceme co nejmenší rozdíly mezi očekávanými a naměřenými hodnotami)

Model nikdy nebude dokonale přesný, kvalitu modelu (good fit) můžeme měřit např. pomocí **indexu determinace** ID^2 (někdy též **koeficient determinace** R^2), který vypočítáme jako podíl sumy čtverců konkrétního modelu a celkové sumy čtverců. Udává, jakou část variability závisle proměnné veličiny Y lze vysvětlit zvolenou regresní funkcí (často v %) a zároveň je mírou těsnosti závislosti proměnné Y na proměnné X.
Vhodnější je model, který má index determinace vyšší.

Testování vhodnosti modelu jako celku (je lepší než náhoda) provádíme pomocí F-testu.

Střední absolutní procentuální chyba predikce MAPE

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right|$$

Vhodnější model je ten, který má hodnotu MAPE nižší.

Předpoklady modelu

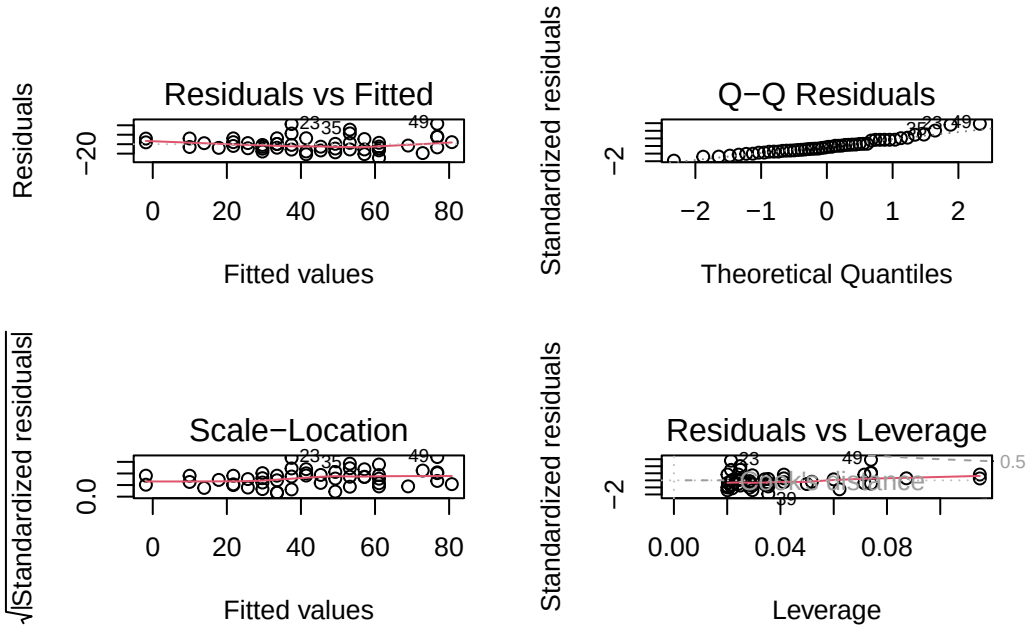
Požadavky na rezidua ověřujeme pomocí analýzy reziduí, chceme, aby rezidua byla:

- nezávislá (Durbin-Watsonův test)
- normálně rozdělená (ověřujeme pomocí kvantil-kvantilového grafu, chceme, aby body ležely na přímce, výpočetně pomocí Shapiro-Wilkova nebo Lillieforsova testu)
- s nulovou střední hodnotou (červená čára vodorovná kolem 0 v grafu Residuals vs. Fitted, tím zároveň ověřujeme linearitu vztahu mezi Y a X, výpočetně t-testem)
- homoskedastická, tj. s konstantním rozptylem (graf Scale-Location, chceme přibližně vodorovnou červenou čáru a rezidua kolem ní rovnoměrně rozmístěna)

Nepřítomnost extrémních odlehlých hodnot (graf Residuals vs. Leverage), linearita vztahu mezi Y a X.

Analýza reziduí graficky

```
model.cars <- lm(dist ~ speed, data = cars)
par(mfrow = c(2,2))
plot(model.cars)
```



Intervaly spolehlivosti a predikční

Pro danou hodnotu x_0 lze počítat meze $100(1-\alpha)\%$ intervalu spolehlivosti pro teoretickou regresní funkci a pro (dostatečně hustou) sekvenci bodů tak dostáváme $100(1-\alpha)\%$ **pás spolehlivosti kolem regresní funkce**. (tj. jde o pás spolehlivosti pro celou přímku)

Analogicky lze vyrobit **predikční pás spolehlivosti**, kdy ovšem jednotlivé predikční intervaly počítáme pro novou hodnotu závislé veličiny Y. (tj, jde o pás spolehlivosti pro jednotlivé nové pozorování)
Interpretace: V jakém rozsahu můžeme očekávat nové pozorování s pravděpodobností alespoň $1-\alpha$.

Pozn.: Predikční interval je vždy širší než interval spolehlivosti.