LINEÁRNÍ REGRESNÍ MODEL

Dvourozměrný náhodný výběr: nezávislé dvojice (,), ..., (,), jsou nezávislé kopie náhodného vektoru (,

*X* a *Y* jsou spojité náhodné veličiny,

např. výška *X* a hmotnost *Y* vysokoškolského studenta.

Data: pozorované (naměřené) číselné hodnoty (,), ..., (,),

například výška a hmotnost *n* studentů.

Sílu lineární závislosti veličin *X* a *Y* měří **korelační koeficient** *ρXY*. Platí | *ρXY* | ≤ 1,

 *ρXY* = 1 <=> Y = , > 0, *ρXY* = -1 <=> Y = , < 0,

*X*, *Y* nezávislé => *X*, *Y* nekorelované ( *ρXY* = 0).

**Výběrový** **korelační koeficient** ,

 = ( - ,

), analogicky.

Teoreticky je náhodná veličina, prakticky číslo vypočítané z dat.

Nabývá hodnot z intervalu ,  **⇒**,

neboli , kde je závisle proměnná,

 je nezávisle proměnná a *ε* je náhodná chyba.

Model **LINEÁRNÍ REGRESE** (regresní přímka)má **předpoklady**:

(i) pro jednotlivé dvojice (,), ..., (,) je

,

(ii) náhodné chyby jsou nezávislé, normálně rozdělené

 s nulovou střední hodnotou a konstantním rozptylem *σ*2,

(iii) jsou nenáhodné.

Z předpokladů plyne:

 **mají podmíněně**

**při známých normální rozdělení**

se střední hodnotou **E**, neboť E

a s rozptylem **var *σ*2**, neboť přičtení konstanty

k náhodné veličině rozptyl neovlivní.

**Práce s modelem regresní přímky**:

- odhad parametrů α

- testování nulovosti parametrů,

- ověření předpokladů,

- posouzení kvality modelu.

**Odhady parametrů *α*, *β* :** **metoda nejmenších čtverců** ...

minimalizujeme součet čtverců odchylek ,...,

bodů o souřadnicích (,), ..., (,)

od proložené přímky , tedy řešíme úlohu

To lze učinit derivováním podle a položením derivací rovných 0.

Řešení: **odhad parametru *β*** je,

 **odhad parametru *α***je .

**Předpovědi**: … odhady pro E

**Rezidua**:

Odhady stejně jako předpovědi a rezidua jsou

teoreticky náhodné veličiny, prakticky čísla spočítaná z dat.

Body (, leží na odhadnuté regresní přímce.

 lze chápat jako předpověď hodnoty veličiny lineárním regresním modelem při známém To se uplatní např. při příchodu nového studenta s výškou jehož hmotnost nemáme možnost změřit.

**Testy hypotéz** o parametrech:

: *α* = 0 ... regresní přímka prochází počátkem,

**: *β* = 0 ... nezávisí formou modelu lineární regrese na** .

 (platí: *β* = 0 <=> *ρXY* = 0.)

zamítáme na hladině 5 %, když:

- je dostatečně velké (odtud testová statistika a kritický obor),

- 0 neleží v 95 % - ním intervalovém odhadu pro *β*,

- p-hodnota je menší než 5 %.

Pro analogicky.

**Analýza reziduí** vypočítaných z dat

znamená grafické **ověření předpokladů**:

- bodové grafy dvojic (), (*i*), by měly mít body

 rovnoměrně roztroušené v rovině kolem nulové úrovně,

 jinak graf () naznačuje nelineární závislost E na ,

 případně nekonstantní rozptyl veličin a ,

 graf (*i*) někdy umožní odhalit závislost mezi veličinami

 a tedy i mezi veličinami

- normalita veličin a :

 histogram reziduí - měl by být jednovrcholový symetrický,

 normální diagram (Q-Q-plot) - měl by mít lineární průběh,

 případně testy normality aplikované na rezidua (p-hodnota > 0,05).

**Koeficient determinace**: nástroj pro posouzení shody modelu s daty.

.

Teoreticky náhodná veličina, prakticky číslo vypočítané z dat.

Nabývá hodnot z intervalu , interpretuje se jako **procento variability závisle proměnné které se podařilo vysvětlit modelem**. V modelu regresní přímky platí: .

**Dobrá shoda modelu s daty** ⇒ předpovědi jsou blízké pozorovaným hodnotám ⇒  **blízký 1**.

Čím menší je , tím spíše působí na ještě jiné vlivy než

nebo se *Y* chová zcela náhodně.

 **MNOHONÁSOBNÁ** **LINEÁRNÍ REGRESE**

Rozšiřuje model regresní přímky na více nezávisle proměnných.

Uvažujme *n* nezávislých kopií náhodných veličin *Y*, :

 , .

Pracujeme s daty, která představuje

pozorovaných realizací závisle proměnné *Y*

a nezávisle proměnných (regresorů)

Např. *Y* … krevní tlak, … věk, … hmotnost,

 … hladina cholesterolu v krvi apod., vše měřeno u *n* osob.

**Předpoklady**:

(i) **,**

(ii ) náhodné chybyjsou nezávislé, normálně rozdělené

 s nulovou střední hodnotou a konstantním rozptylem *σ*2,

(iii) jsou nenáhodné,

(iv) matice má lineárně nezávislé sloupce.

*Y* je spojitá náhodná veličina, regresory většinou spojité, ojediněle diskrétní. Např. … pohlaví (0 muž, 1 žena)

=> model má pro muže konstantu a pro ženy .

Z předpokladů plyne: náhodné veličiny

,

**podmíněně při známých normálně rozdělené**,

**, var *σ*2**.

**Odhady parametrů**:metoda nejmenších čtverců ...

.

Řešení lze spočítat maticově: = .

**Testy hypotéz** o parametrech:

: = 0 ... model neobsahuje konstantu (intercept),

 **= 0 ... *Y* nezávisí formou modelu**

 **mnohonásobné lineární regrese na , , .**

**Předpovědi** :

**Rezidua**:

 jsou odhadem středních hodnot , reprezentují předpověď hodnoty veličiny modelem.

**Analýza reziduí** =ověření předpokladů

probíhá analogicky jako v modelu regresní přímky, k odhalení

nelinearity závislosti E na a nekonstantního rozptylu a

se kreslí graf dvojic (),

**Normalita náhodných chyb** a tedy i veličin

není potřeba pro odhad parametrů modelu metodou nejmenších čtverců, ale **je potřeba pro testování hypotéz o parametrech**

k určení kritických oborů, intervalových odhadůparametrů

ap-hodnot.

**Nesplnění normality** reziduí: výsledky **testování nulovosti parametrů**  **lze** **zkontrolovat pomocí koeficientu determinace** (počítá se stejně jako v modelu regresní přímky):

malé snížení koeficientu determinace v modelu s vyloučeným regresorem oproti původnímu modelu ⇒v modelu mnohonásobné lineární regrese nepřispívá zásadně k vysvětlení variability závisle proměnné *Y* ( lze považovat za nulový).