Regresné modely pre nespojité veličiny

I. Žežula

Prírodovedecká fakulta
Univerzity P. J. Šafárika, Košice

18. slovenská štatistická konferencia, Košice 2016
23.6. – 25.6.2016

Obsah

1. Logistická regresia
   - Úvod
   - Základný model
   - Všeobecnejšie prediktory
   - Všeobecný model
   - Testy asociácie
2. Multinomická regresia
   - Úvod
   - Všeobecný model
   - Miery dobrej zhody a prebytok rozptylu
   - Testy
3. Ordinálna regresia
   - Úvod
   - Model proporcionálnych šancí
   - Ďalšie modely
4. Poissonova regresia
   - Úvod
   - Základný model
   - Všeobecný model

1) Logistická regresia

- **Úvod**

Testovanie rizikového faktora: chceme overiť, či určitý faktor má vplyv na pravdepodobnosť prepuknutia nejakej sledovanej choroby.

Tomu zodpovedá nasledujúca kontingenčná tabuľka:

rizikový faktor	zdravotný stav		suma
	chorý	zdravý	
prítomný	n_{11}	n_{12}	n_{10}
neprítomný	n_{21}	n_{22}	n_{20}
suma	n_{01}	n_{02}	n

1) Logistická regresia

- **Úvod**

Testovanie rizikového faktora: chceme overiť, či určitý faktor má vplyv na pravdepodobnosť prepuknutia nejakej sledovanej choroby.

Tomu zodpovedá nasledujúca kontingenčná tabuľka:

rizikový faktor	zdravotný stav		suma
	chorý	zdravý	
prítomný	n_{11}	n_{12}	n_{10}
neprítomný	n_{21}	n_{22}	n_{20}
suma	n_{01}	n_{02}	n

- Šance prepuknutia choroby pre obidve skupiny sú:

$$o_1 = \frac{n_{11}}{n_{12}} = \frac{n_{11}}{n_{10} - n_{11}} = \frac{\frac{n_{11}}{n_{10}}}{1 - \frac{n_{11}}{n_{10}}} = \frac{\hat{p}_1}{1 - \hat{p}_1}, \qquad o_2 = \frac{\hat{p}_2}{1 - \hat{p}_2}$$

1) Logistická regresia

- Ako mieru rizika môžeme vziať **pomer šancí** (odds ratio):

$$\mathbf{OR} = \frac{o_1}{o_2} = \frac{\hat{p}_1}{1-\hat{p}_1} \cdot \frac{1-\hat{p}_2}{\hat{p}_2}$$

$$o_1 = \mathbf{OR} \cdot o_2$$

1) Logistická regresia

- Ako mieru rizika môžeme vziať **pomer šancí** (odds ratio):

$$\mathbf{OR} = \frac{o_1}{o_2} = \frac{\hat{p}_1}{1-\hat{p}_1} \cdot \frac{1-\hat{p}_2}{\hat{p}_2}$$
$$o_1 = \mathbf{OR} \cdot o_2$$

- Potom platí

$$\log(o_1) = \log(o_2) + \log(\mathbf{OR}),$$

1) Logistická regresia

- Ako mieru rizika môžeme vziať **pomer šancí** (odds ratio):

$$\mathbf{OR} = \frac{o_1}{o_2} = \frac{\hat{p}_1}{1-\hat{p}_1} \cdot \frac{1-\hat{p}_2}{\hat{p}_2}$$
$$o_1 = \mathbf{OR} \cdot o_2$$

- Potom platí

$$\log(o_1) = \log(o_2) + \log(\mathbf{OR}),$$

- čiže

$$y = \log(o_2) + \log(\mathbf{OR}) \cdot x,$$

kde $x \in \{0, 1\}$ a $y \in \{\log(o_1), \log(o_2)\}$.

1) Logistická regresia

- Toto je základný logistický model. Formálne je to regresný model

$$y = \beta_0 + \beta_1 x$$

so základnou úrovňou $\beta_0 = \log(o_2)$ a smernicou $\beta_1 = \log(\mathbf{OR})$ – efektom prítomnosti rizikového faktora.

1) Logistická regresia

- Toto je základný logistický model. Formálne je to regresný model

$$y = \beta_0 + \beta_1 x$$

so základnou úrovňou $\beta_0 = \log(o_2)$ a smernicou $\beta_1 = \log(\mathbf{OR})$ – efektom prítomnosti rizikového faktora.

- Ak označíme pravdepodobnosť nastatia udalosti (prepuknutia choroby) p, potom

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

v obidvoch skupinách.

1) Logistická regresia

- Toto je základný logistický model. Formálne je to regresný model

$$y = \beta_0 + \beta_1 x$$

so základnou úrovňou $\beta_0 = \log(o_2)$ a smernicou $\beta_1 = \log(\mathbf{OR})$ – efektom prítomnosti rizikového faktora.

- Ak označíme pravdepodobnosť nastatia udalosti (prepuknutia choroby) p, potom

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

v obidvoch skupinách.

- Z toho

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

1) Logistická regresia

- Odhadovanie neznámych parametrov sa robí metódou maximálnej vierohodnoti (ML-metódou). Vierohodnosť pozorovaných početností pre dané β-y je

$$L(\beta) = \binom{n_{10}}{n_{11}} p_1^{n_{11}} (1-p_1)^{n_{10}-n_{11}} \binom{n_{20}}{n_{21}} p_2^{n_{21}} (1-p_2)^{n_{20}-n_{21}}$$

1) Logistická regresia

- Odhadovanie neznámych parametrov sa robí metódou maximálnej vierohodnoti (ML-metódou). Vierohodnosť pozorovaných početností pre dané β-y je

$$L(\beta) = \binom{n_{10}}{n_{11}} p_1^{n_{11}} (1-p_1)^{n_{10}-n_{11}} \binom{n_{20}}{n_{21}} p_2^{n_{21}} (1-p_2)^{n_{20}-n_{21}}$$

- Z toho

$$\ell(\beta) = \log L(\beta) = n_{11} \log(p_1) + n_{12} \log(1-p_1) + \\ + n_{21} \log(p_2) + n_{22} \log(1-p_2)$$

1) Logistická regresia

- Odhadovanie neznámych parametrov sa robí metódou maximálnej vierohodnoti (ML-metódou). Vierohodnosť pozorovaných početností pre dané β-y je

$$L(\beta) = \binom{n_{10}}{n_{11}} p_1^{n_{11}} (1-p_1)^{n_{10}-n_{11}} \binom{n_{20}}{n_{21}} p_2^{n_{21}} (1-p_2)^{n_{20}-n_{21}}$$

- Z toho

$$\ell(\beta) = \log L(\beta) = n_{11} \log(p_1) + n_{12} \log(1-p_1) + \\ + n_{21} \log(p_2) + n_{22} \log(1-p_2)$$

- Ľahko sa ukáže, že ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta_1} = n_{11} - n_{10} p_1 = 0\,, \quad \frac{\partial \ell}{\partial \beta_0} = n_{11} - n_{10} p_1 + n_{21} - n_{20} p_2 = 0$$

1) Logistická regresia

- V tomto prípade už riešenie poznáme, $\hat{p}_1 = \frac{n_{11}}{n_{10}}$ a $\hat{p}_2 = \frac{n_{21}}{n_{20}}$ (a príslušné β-y).

1) Logistická regresia

- V tomto prípade už riešenie poznáme, $\hat{p}_1 = \frac{n_{11}}{n_{10}}$ a $\hat{p}_2 = \frac{n_{21}}{n_{20}}$ (a príslušné β-y).
- Vieme odvodiť aj asymptotickú variančnú maticu $\hat{\beta}$. Jej tvar je

$$\text{var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = J^{-1} = \frac{1}{ab}\begin{pmatrix} a & -a\\ -a & a+b\end{pmatrix},$$

kde $a = n_{10}\hat{p}_1\,(1-\hat{p}_1)$, $b = n_{20}\hat{p}_2\,(1-\hat{p}_2)$.

1) Logistická regresia

- V tomto prípade už riešenie poznáme, $\hat{p}_1 = \frac{n_{11}}{n_{10}}$ a $\hat{p}_2 = \frac{n_{21}}{n_{20}}$ (a príslušné β-y).
- Vieme odvodiť aj asymptotickú variančnú maticu $\hat{\beta}$. Jej tvar je

$$\text{var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = J^{-1} = \frac{1}{ab}\begin{pmatrix} a & -a\\ -a & a+b\end{pmatrix},$$

  kde $a = n_{10}\hat{p}_1(1-\hat{p}_1)$, $b = n_{20}\hat{p}_2(1-\hat{p}_2)$.
- $J = \begin{pmatrix} a+b & a\\ a & a\end{pmatrix}$ je informačná matica.

1) Logistická regresia

- **Príklad:** Baystate Medical Center v Springfielde (MA), USA, študoval faktory ovplyvňujúce nízku pôrodnú hmotnosť novorodencov. Vezmime za rizikový faktor fajčenie v priebehu tehotenstva. Dostaneme nasledujúcu kontingenčnú tabuľku:

fajčenie	nízka pôrodná hmotnosť		suma
	áno	nie	
prítomné	30	44	74
neprítomné	29	86	115
suma	59	130	159

1) Logistická regresia

- **Príklad:** Baystate Medical Center v Springfielde (MA), USA, študoval faktory ovplyvňujúce nízku pôrodnú hmotnosť novorodencov. Vezmime za rizikový faktor fajčenie v priebehu tehotenstva. Dostaneme nasledujúcu kontingenčnú tabuľku:

fajčenie	nízka pôrodná hmotnosť		suma
	áno	nie	
prítomné	30	44	74
neprítomné	29	86	115
suma	59	130	159

- Šanca pri fajčiarkách je $o_1 = 30/44 = 0.681818$, pri nefajčiarkách $o_2 = 29/86 = 0.337209$, **OR** $= 2.021644$. Sakodov koeficient $S = 0.225344$ indikuje miernu, ale štatisticky významnú závislosť ($\chi^2 = 4.92 > 3.84 = \chi^2_1(0.05)$).

1) Logistická regresia

- Keďže $\log(2.021644) = 0.7040592$, $\log(0.337209) = -1.087051$, dostávame model

$$y = -1.087 + 0.704\,x.$$

1) Logistická regresia

- Keďže $\log(2.021644) = 0.7040592$, $\log(0.337209) = -1.087051$, dostávame model

$$y = -1.087 + 0.704\,x.$$

- Výstup z komerčného softwaru:

	coefficient	*std. error*	*z*	$P > \|z\|$	*95% conf. interval*	
fajcenie	*0.7040592*	*0.319639*	*2.20*	*0.028*	*0.077579*	*1.330539*
constant	*-1.087051*	*0.21473*	*-5.06*	*0.000*	*-1.50791*	*-0.66619*

1) Logistická regresia

- Keďže $\log(2.021644) = 0.7040592$, $\log(0.337209) = -1.087051$, dostávame model

$$y = -1.087 + 0.704\,x.$$

- Výstup z komerčného softwaru:

	coefficient	*std. error*	*z*	$P > \|z\|$	*95% conf. interval*	
fajcenie	*0.7040592*	*0.319639*	*2.20*	*0.028*	*0.077579*	*1.330539*
constant	*-1.087051*	*0.21473*	*-5.06*	*0.000*	*-1.50791*	*-0.66619*

- Vieme získať aj 95% interval spoľahlivosti (IS) pre **OR**:

$$\left[e^{0.077579}; e^{1.330539}\right] = [1.08;\ 3.78]$$

1) Logistická regresia

Všeobecný kategoriálny prediktor:

Musíme odhadovať viac pravdepodobností. Nech počet kategórií prediktora je m:

- $o_i = n_{i1}/n_{i2}\,,\ i = 1, \ldots, m$

1) Logistická regresia

Všeobecný kategoriálny prediktor:

Musíme odhadovať viac pravdepodobností. Nech počet kategórií prediktora je m:

- $o_i = n_{i1}/n_{i2}\,,\ i = 1, \ldots, m$
- $\log(\mathbf{OR}_i) = \log \dfrac{o_i}{o_m} = \beta_i,\ i = 1, \ldots, m-1$

1) Logistická regresia

Všeobecný kategoriálny prediktor:

Musíme odhadovať viac pravdepodobností. Nech počet kategórií prediktora je m:

- $o_i = n_{i1}/n_{i2}\,,\ i = 1, \ldots, m$
- $\log(\mathbf{OR}_i) = \log \dfrac{o_i}{o_m} = \beta_i,\ i = 1, \ldots, m-1$

Potom

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{m-1} x_{m-1}\,,$$

kde všetky $x_i \in \{0, 1\}$, ale iba jedno z nich nadobúda hodnotu 1.

1) Logistická regresia

Všeobecný kategoriálny prediktor:

Musíme odhadovať viac pravdepodobností. Nech počet kategórií prediktora je m:

- $o_i = n_{i1}/n_{i2}\,,\ i = 1, \ldots, m$
- $\log(\mathbf{OR}_i) = \log \dfrac{o_i}{o_m} = \beta_i,\ i = 1, \ldots, m-1$

Potom

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{m-1} x_{m-1}\,,$$

kde všetky $x_i \in \{0, 1\}$, ale iba jedno z nich nadobúda hodnotu 1. Jedna kategória z m musí byť referenčnou kategóriou. Jedna vysvetľujúca premenná sa nahradí $m-1$ indikátormi ostatných kategórií, ktoré sa navzájom vylučujú. Významnosť regresného koeficientu indikuje významný rozdiel medzi príslušnou kategóriou a referenčnou kategóriou (diferenciálny efekt kategórie).

1) Logistická regresia

Príklad: Vezmime hmotnosť matky (zoskupenú do 3 kategórií) ako rizikový faktor nízkej pôrodnej hmotnosti novorodenca. Dostaneme

hmotnostná skupina [lb]	*nízka pôrodná hmotnosť*		*suma*	*riadkové podiely*	
	áno	*nie*			
≤ 110	*25*	*28*	*53*	*47.2%*	*52.8%*
$(110; 150]$	*27*	*73*	*100*	*27.0%*	*73.0%*
> 150	*7*	*29*	*36*	*19.4%*	*80.6%*
suma	*59*	*130*	*189*	*31.2%*	*68.8%*

Meniace sa riadkové podiely ukazujú, že (nízka) hmotnosť matky môže byť rizikovým faktorom.

1) Logistická regresia

Výstup zo softwaru pre 3. kategóriu ako referenčnú:

variable	*coefficient*	*std. error*	*Wald*	*df*	*p-value*
wt_groups			*9.073891*	*2*	*0.010706*
wt_groups(1)	*1.308056995*	*0.503044916*	*6.761449*	*1*	*0.009315*
wt_groups(2)	*0.426763105*	*0.477572579*	*0.798537*	*1*	*0.371531*
constant	*-1.42138568*	*0.421117444*	*11.39246*	*1*	*0.000737*

Výstup zo softwaru pre 1. kategóriu ako referenčnú:

variable	*coefficient*	*std. error*	*Wald*	*df*	*p-value*
wt_groups			*9.073891*	*2*	*0.010706*
wt_groups(2)	*-0.88129389*	*0.355598021*	*6.142184*	*1*	*0.013199*
wt_groups(3)	*-1.308057*	*0.503044916*	*6.761449*	*1*	*0.009315*
constant	*-0.11332869*	*0.27516229*	*0.16963*	*1*	*0.680441*

Symbolicky zapísané, pre vplyv hmotnostných skupín platí $1 \neq \{2, 3\}$ na hladine 5%. Hmotnosť matky je štatisticky významný faktor.

1) Logistická regresia

Kvantitatívny prediktor:

Ak máme kvantitatívnu vysvetľujúcu premennú, ktorá ovplyvňuje pravdepodobnosť výsledku, môžeme jednoducho predpokladať, že p je (spojitou) funkciou x:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

1) Logistická regresia

Kvantitatívny prediktor:

Ak máme kvantitatívnu vysvetľujúcu premennú, ktorá ovplyvňuje pravdepodobnosť výsledku, môžeme jednoducho predpokladať, že p je (spojitou) funkciou x:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Regresný koeficient x sa interpretuje ako efekt jednotkovej zmeny x na výsledok.

1) Logistická regresia

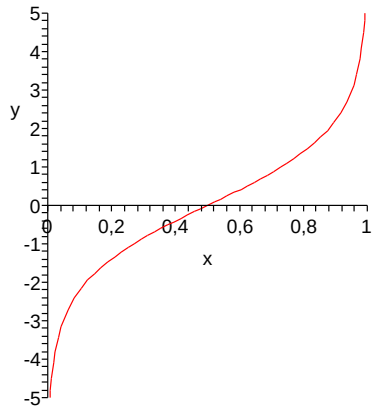
Kvantitatívny prediktor:

Ak máme kvantitatívnu vysvetľujúcu premennú, ktorá ovplyvňuje pravdepodobnosť výsledku, môžeme jednoducho predpokladať, že p je (spojitou) funkciou x:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Regresný koeficient x sa interpretuje ako efekt jednotkovej zmeny x na výsledok.

Logistická transformácia $p \to \log\left(\frac{p}{1-p}\right)$ je z $(0;1)$ na $(-\infty;+\infty)$, takže odstraňuje ohraničenia obmedzujúce regresiu.



1) Logistická regresia

Príklad: Vezmime hmotnosť matky ako spojitý rizikový faktor nízkej pôrodnej hmotnosti novorodenca.
Software dáva:

variable	*coefficient*	*std. error*	*Wald*	*df*	*p-value*
lwt	*-0.01405826*	*0.006169588*	*5.192193*	*1*	*0.022689*
constant	*0.998314313*	*0.78529092*	*1.616119*	*1*	*0.203634*

Hmotnosť matky je opäť štatisticky významný faktor.

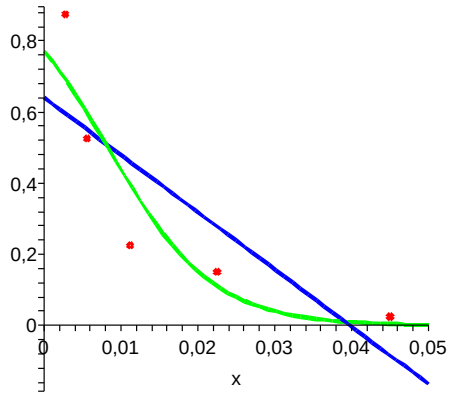
1) Logistická regresia

Príklad: Vezmime hmotnosť matky ako spojitý rizikový faktor nízkej pôrodnej hmotnosti novorodenca.
Software dáva:

variable	*coefficient*	*std. error*	*Wald*	*df*	*p-value*
lwt	*-0.01405826*	*0.006169588*	*5.192193*	*1*	*0.022689*
constant	*0.998314313*	*0.78529092*	*1.616119*	*1*	*0.203634*

Hmotnosť matky je opäť štatisticky významný faktor.

Príklad: Bol skúmaný efekt anti-pneumokokového séra na prežitie chorých myší. Bolo podaných 5 rôznych dávok séra piatim skupinám po 40 myší. Graf ukazuje percentuálnu úmrtnosť, jednoduchú regresnú priamku a logistickú regresnú krivku.



1) Logistická regresia

Všeobecný logistický regresný model:

- Máme dichotomickú závislú premennú Y a vysvetľujúce premenné $X_1, X_2, \ldots, X_k$ akéhokoľvek typu. Chceme vysvetliť alebo predpovedať správanie Y pomocou vysvetľujúcich premenných.

1) Logistická regresia

Všeobecný logistický regresný model:

- Máme dichotomickú závislú premennú Y a vysvetľujúce premenné $X_1, X_2, \ldots, X_k$ akéhokoľvek typu. Chceme vysvetliť alebo predpovedať správanie Y pomocou vysvetľujúcich premenných.
- Model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

alebo (položiac $X_0 = \mathbf{1}$)

$$p_i = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)} = \frac{1}{1 + \exp\left(-\sum_{j=0}^{k} \beta_j x_{ij}\right)}$$

1) Logistická regresia

Všeobecný logistický regresný model:

- Máme dichotomickú závislú premennú Y a vysvetľujúce premenné $X_1, X_2, \ldots, X_k$ akéhokoľvek typu. Chceme vysvetliť alebo predpovedať správanie Y pomocou vysvetľujúcich premenných.
- Model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

alebo (položiac $X_0 = \mathbf{1}$)

$$p_i = \frac{\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)} = \frac{1}{1 + \exp\left(-\sum_{j=0}^{k} \beta_j x_{ij}\right)}$$

- Odhadovanie parametrov sa robí pomocou ML-metódy.

1) Logistická regresia

- Vierohodnosť pozorovaných početností pre dané β-y je

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

1) Logistická regresia

- Vierohodnosť pozorovaných početností pre dané β-y je

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

- Z toho

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

1) Logistická regresia

- Vierohodnosť pozorovaných početností pre dané β-y je

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1-y_i}$$

- Z toho

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n} y_i \log\left(p_i\right) + \left(1 - y_i\right) \log\left(1 - p_i\right)$$

- ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \left(y_i - p_i(\beta)\right) x_{ij} = 0 \quad \forall j$$

1) Logistická regresia

- Tieto rovnice sa vo všeobecnosti riešia numericky (Newtonov-Raphsonov typ algoritmu).

1) Logistická regresia

- Tieto rovnice sa vo všeobecnosti riešia numericky (Newtonov-Raphsonov typ algoritmu).
- Označme $X = \{x_{ij}\}$, $V = \text{diag}\{\hat{p}_1, \ldots, \hat{p}_n\}$. Informačná matica je

$$J = X'VX$$

a J^{-1} je asymptotická variančná matica $\hat{\beta}$.

1) Logistická regresia

- Tieto rovnice sa vo všeobecnosti riešia numericky (Newtonov-Raphsonov typ algoritmu).
- Označme $X = \{x_{ij}\}$, $V = \text{diag}\{\hat{p}_1, \ldots, \hat{p}_n\}$. Informačná matica je

$$J = X'VX$$

  a J^{-1} je asymptotická variančná matica $\hat{\beta}$.
- Ak všetky prediktory sú kategoriálne, dá sa model preformulovať pre binomické premenné.

1) Logistická regresia

- Tieto rovnice sa vo všeobecnosti riešia numericky (Newtonov-Raphsonov typ algoritmu).
- Označme $X = \{x_{ij}\}$, $V = \text{diag}\{\hat{p}_1, \ldots, \hat{p}_n\}$. Informačná matica je

$$J = X'VX$$

  a J^{-1} je asymptotická variančná matica $\hat{\beta}$.
- Ak všetky prediktory sú kategoriálne, dá sa model preformulovať pre binomické premenné.
- Klady & zápory:

1) Logistická regresia

- Tieto rovnice sa vo všeobecnosti riešia numericky (Newtonov-Raphsonov typ algoritmu).
- Označme $X = \{x_{ij}\}$, $V = \text{diag}\{\hat{p}_1, \ldots, \hat{p}_n\}$. Informačná matica je

$$J = X'VX$$

  a J^{-1} je asymptotická variančná matica $\hat{\beta}$.
- Ak všetky prediktory sú kategoriálne, dá sa model preformulovať pre binomické premenné.
- Klady & zápory:
  - nemáme explicitné vzorce, odhady sú iteratívne

1) Logistická regresia

- Tieto rovnice sa vo všeobecnosti riešia numericky (Newtonov-Raphsonov typ algoritmu).
- Označme $X = \{x_{ij}\}$, $V = \text{diag}\{\hat{p}_1, \ldots, \hat{p}_n\}$. Informačná matica je

$$J = X'VX$$

  a J^{-1} je asymptotická variančná matica $\hat{\beta}$.
- Ak všetky prediktory sú kategoriálne, dá sa model preformulovať pre binomické premenné.
- Klady & zápory:
  - – nemáme explicitné vzorce, odhady sú iteratívne
  - \+ sú k dispozícii približné variancie, p-hodnoty a IS

1) Logistická regresia

- Waldov test koeficientu:
  Ak platí $H_0 : \beta_i = 0$, potom

$$Z = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

  má asymptoticky rozdelenie $N(0; 1)$.
  Existuje alternatívna chí-kvadrát forma (Z^2).

1) Logistická regresia

- Waldov test koeficientu:
  Ak platí $H_0 : \beta_i = 0$, potom

$$Z = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

  má asymptoticky rozdelenie $N(0; 1)$.
  Existuje alternatívna chí-kvadrát forma (Z^2).
- Test pomerom vierohodností (LRT):
  Pojmy:
  - odhadovaný model – model s prediktormi
  - prázdny model – model bez prediktorov, iba s konštantou
  - plný model – model, ktorý predpovedá $n_{i0}\hat{p}_i = n_{i1}\ \forall i$, t.j. $\hat{y}_i = y_i\ \forall i$

1) Logistická regresia

Deviancia:

$$\hat{\ell}_m = \log \hat{L}_m = \log \hat{L}(\text{odhadovaný model}),$$
$$\hat{\ell}_f = \log \hat{L}_f = \log \hat{L}(\text{plný model})$$
$$D_m = 2\left(\hat{\ell}_f - \hat{\ell}_m\right) = 2\log\left(\hat{L}_f/\hat{L}_m\right)$$

Pre binomické dáta je D_m mierou dobrej zhody dát s modelom.
Asymptoticky

$$D_m \approx \chi^2_{n-k-1}$$

1) Logistická regresia

Deviancia:

$$\hat{\ell}_m = \log \hat{L}_m = \log \hat{L}(\text{odhadovaný model}),$$
$$\hat{\ell}_f = \log \hat{L}_f = \log \hat{L}(\text{plný model})$$
$$D_m = 2\left(\hat{\ell}_f - \hat{\ell}_m\right) = 2\log\left(\hat{L}_f/\hat{L}_m\right)$$

Pre binomické dáta je D_m mierou dobrej zhody dát s modelom. Asymptoticky

$$D_m \approx \chi^2_{n-k-1}$$

Pre všetky druhy modelov platí, že rozdiel deviancií sa dá použiť na porovnanie vnorených modelov (LRT test významnosti pridaných prediktorov):

$$D_{m_1} - D_{m_2} = 2\left(\hat{\ell}_{m_2} - \hat{\ell}_{m_1}\right) \approx \chi^2_{k_2-k_1}$$

1) Logistická regresia

Príklad: Uvažujme opäť pôrodné dáta s rizikovým faktorom fajčenie. Software dáva

$$\hat{\ell}_1 = \log \hat{L}_1 = \log \hat{L}(\text{odhadovaný model}) = -114.9023,$$
$$\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}(\text{prázdny model}) = -117.336.$$

Ak platí $H_0 : \beta_1 = 0$, potom

$$2\left(\hat{\ell}_1 - \hat{\ell}_0\right) = 2\log\left(\frac{\hat{L}_1}{\hat{L}_0}\right) \approx \chi^2_1$$

Máme teda

$$\Delta D = 2(-114.9023 + 117.336) \;=\; 4.8674 > 3.84 = \chi^2_1(0.05),$$

takže asociácia medzi fajčením a nízkou pôrodnou hmotnosťou je významná.

1) Logistická regresia

Interakcie

Logistický regresný model umožňuje uvažovať (a testovať) aj interakcie kategoriálnych premenných. Ak premenná X má c kategórií a premenná Z d kategórií, potom interakcia $X * Z$ má $(c-1)(d-1)$ kategórií. Sú to všetky možné kombinácie nereferenčných kategórií X a Z.

1) Logistická regresia

Interakcie

Logistický regresný model umožňuje uvažovať (a testovať) aj interakcie kategoriálnych premenných. Ak premenná X má c kategórií a premenná Z d kategórií, potom interakcia $X * Z$ má $(c-1)(d-1)$ kategórií. Sú to všetky možné kombinácie nereferenčných kategórií X a Z.

Interakcie sa testujú Waldovým testom alebo LRT testom ako akákoľvek iná premenná.

1) Logistická regresia

Interakcie

Logistický regresný model umožňuje uvažovať (a testovať) aj interakcie kategoriálnych premenných. Ak premenná X má c kategórií a premenná Z d kategórií, potom interakcia $X * Z$ má $(c-1)(d-1)$ kategórií. Sú to všetky možné kombinácie nereferenčných kategórií X a Z.

Interakcie sa testujú Waldovým testom alebo LRT testom ako akákoľvek iná premenná.

Ak je interakcia významná, významnosť pôvodných interagujúcich premenných nemá interpretáciu any zmysel. Efekty sú skrížené, a preto máme len dve možnosti riešenia:

1. Urobiť stratifikáciu a rôzne skupiny/vrstvy analyzovať zvlášť.
2. Vytvoriť novú premennú, ktorej obor hodnôt je kartézsky súčin skrížených premenných (interakčná premenná), a vynechať interagujúce premenné.

1) Logistická regresia

Miery dobrej zhody

Nech $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}$(prázdny model).

- McFaddenovo $R^2_{MF} = 1 - \dfrac{\hat{\ell}_m}{\hat{\ell}_0}$

1) Logistická regresia

Miery dobrej zhody

Nech $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}$(prázdny model).

- McFaddenovo $R^2_{MF} = 1 - \dfrac{\hat{\ell}_m}{\hat{\ell}_0}$
- Coxovo & Snellovo $R^2_{CS} = 1 - \left(\dfrac{\hat{L}_0}{\hat{L}_m}\right)^{\frac{2}{n}}$

1) Logistická regresia

Miery dobrej zhody

Nech $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}$(prázdny model).

- McFaddenovo $R^2_{MF} = 1 - \dfrac{\hat{\ell}_m}{\hat{\ell}_0}$
- Coxovo & Snellovo $R^2_{CS} = 1 - \left(\dfrac{\hat{L}_0}{\hat{L}_m}\right)^{\frac{2}{n}}$
- Nagelkerkovo $R^2_N = \dfrac{R^2_{CS}}{1 - \hat{L}_0^{\frac{2}{n}}}$

1) Logistická regresia

Miery dobrej zhody

Nech $\hat{\ell}_0 = \log \hat{L}_0 = \log \hat{L}$(prázdny model).

- McFaddenovo $R^2_{MF} = 1 - \frac{\hat{\ell}_m}{\hat{\ell}_0}$
- Coxovo & Snellovo $R^2_{CS} = 1 - \left(\frac{\hat{L}_0}{\hat{L}_m}\right)^{\frac{2}{n}}$
- Nagelkerkovo $R^2_N = \frac{R^2_{CS}}{1 - \hat{L}_0^{\frac{2}{n}}}$
- Hosmer-Lemeshowov test (chí-kvadrát test dobrej zhody v kontingenčnej tabuľke medzi závislou premennou a skupinami predikovaných hodnôt)

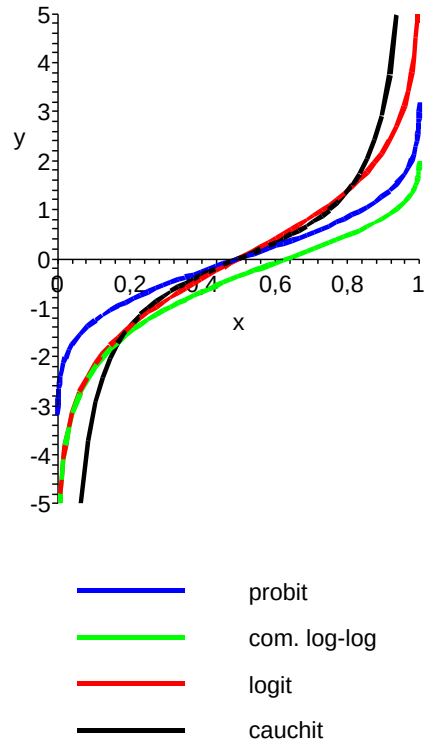
1) Logistická regresia

Alternatívy

Logistická funkcia nie je jediná, ktorú možno použiť na transformáciu pravdepodobností binárnych výsledkov. Najpoužívanejšie sú:

- logistická funkcia $\log\left(\frac{p}{1-p}\right)$
- probitová funkcia $\Phi^{-1}(p)$
- komplementárna log-log funkcia $\log(-\log(1-p))$
- negatívna log-log funkcia $-\log(-\log(p))$
- cauchitová funkcia $\tan\left(\left(p-\frac{1}{2}\right)\pi\right)$

Nazývajú sa spojovacie funkcie.



2) Multinomická regresia

- **Príklad:** Z dát 1991 U.S. General Social Survey chceme overiť, či pohlavie respondenta ovplyvňuje pravdepodobnosť úrovne pocitu spokojnosti so životom.

  Dostávame nasledujúcu kontingenčnú tabuľku:

pohlavie	úroveň spokojnosti			suma
	vzrušujúci	rutinný	nudný	
muži	213	200	12	425
ženy	221	305	29	555
suma	434	505	41	980

2) Multinomická regresia

- **Príklad:** Z dát 1991 U.S. General Social Survey chceme overiť, či pohlavie respondenta ovplyvňuje pravdepodobnosť úrovne pocitu spokojnosti so životom.

  Dostávame nasledujúcu kontingenčnú tabuľku:

pohlavie	úroveň spokojnosti			suma
	vzrušujúci	rutinný	nudný	
muži	213	200	12	425
ženy	221	305	29	555
suma	434	505	41	980

- Toto by sa dalo riešiť pomocou série binárnych modelov:

2) Multinomická regresia

- **Príklad:** Z dát 1991 U.S. General Social Survey chceme overiť, či pohlavie respondenta ovplyvňuje pravdepodobnosť úrovne pocitu spokojnosti so životom.

  Dostávame nasledujúcu kontingenčnú tabuľku:

pohlavie	úroveň spokojnosti			suma
	vzrušujúci	rutinný	nudný	
muži	213	200	12	425
ženy	221	305	29	555
suma	434	505	41	980

- Toto by sa dalo riešiť pomocou série binárnych modelov:
  1. život je vzrušujúci – nie vzrušujúci
  2. nie vzrušujúci život: rutinný – nudný

2) Multinomická regresia

- Alternatívou je uvažovať viac pravdepodobností a šancí. V našom príklade to znamená uvažovať dve multinomické rozdelenia (p_{11}, p_{12}, p_{13}) a (p_{21}, p_{22}, p_{23}), ktoré popisujú pravdepodobnosti úrovní pocitu spokojnosti zvlášť pre pre mužov a ženy. Najjednoduchší spôsob je zvoliť si jednu z kategórií ako referenčnú – povedzme vzrušujúci život – lebo jedna z pravdepodobností v každom riadku je redundantná.

2) Multinomická regresia

- Alternatívou je uvažovať viac pravdepodobností a šancí. V našom príklade to znamená uvažovať dve multinomické rozdelenia (p_{11}, p_{12}, p_{13}) a (p_{21}, p_{22}, p_{23}), ktoré popisujú pravdepodobnosti úrovní pocitu spokojnosti zvlášť pre pre mužov a ženy. Najjednoduchší spôsob je zvoliť si jednu z kategórií ako referenčnú – povedzme vzrušujúci život – lebo jedna z pravdepodobností v každom riadku je redundantná.
- Výsledný model teda je

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) = \beta_{0j} + \beta_{1j}x_i\,, \quad j = 2, 3,$$

kde $x_i \in \{0, 1\}$ je indikátor pohlavia.

2) Multinomická regresia

- Alternatívou je uvažovať viac pravdepodobností a šancí. V našom príklade to znamená uvažovať dve multinomické rozdelenia (p_{11}, p_{12}, p_{13}) a (p_{21}, p_{22}, p_{23}), ktoré popisujú pravdepodobnosti úrovní pocitu spokojnosti zvlášť pre pre mužov a ženy. Najjednoduchší spôsob je zvoliť si jednu z kategórií ako referenčnú – povedzme vzrušujúci život – lebo jedna z pravdepodobností v každom riadku je redundantná.
- Výsledný model teda je

$$\log\left(\frac{p_{ij}}{p_{i1}}\right) = \beta_{0j} + \beta_{1j}x_i\,, \quad j = 2, 3,$$

  kde $x_i \in \{0, 1\}$ je indikátor pohlavia.
- Predchádzajúci vzorec je totožný s jednoduchým lineárnym logistickým modelom v prípade dichotomických výsledkov.

2) Multinomická regresia

Na základe pozorovaných početností dostávame nasledujúce šance a ich logaritmy:

šanca		*log-šanca*	
$200/213 = 0.938967$	$12/213 = 0.056338$	*-0.06297*	*-2.87639*
$305/221 = 1.38009$	$29/221 = 0.131222$	*0.322149*	*-2.03087*

2) Multinomická regresia

Na základe pozorovaných početností dostávame nasledujúce šance a ich logaritmy:

šanca		*log-šanca*	
$200/213 = 0.938967$	$12/213 = 0.056338$	*-0.06297*	*-2.87639*
$305/221 = 1.38009$	$29/221 = 0.131222$	*0.322149*	*-2.03087*

Rozdiely logaritmov šancí v posledných dvoch stĺpcoch sú -0.385123874 a -0.845518644. Môžeme teda napísať dva modely

$$y = 0.322 - 0.385x, \quad y = -2.031 - 0.846x$$

pre pocit rutiny a nudy.

2) Multinomická regresia

Na základe pozorovaných početností dostávame nasledujúce šance a ich logaritmy:

šanca		*log-šanca*	
$200/213 = 0.938967$	$12/213 = 0.056338$	*-0.06297*	*-2.87639*
$305/221 = 1.38009$	$29/221 = 0.131222$	*0.322149*	*-2.03087*

Rozdiely logaritmov šancí v posledných dvoch stĺpcoch sú -0.385123874 a -0.845518644. Môžeme teda napísať dva modely

$$y = 0.322 - 0.385x, \quad y = -2.031 - 0.846x$$

pre pocit rutiny a nudy. Software dáva:

Life		B	Std. Error	Wald	df	Sig.	Exp(B)	95% CI	
Routine	Intercept	0.322149	0.088338	13.29904	1	0.000266			
	[sex=1]	-0.38512	0.132282	8.476221	1	0.003598	0.680366	0.524982	0.881741
	[sex=2]	0			0				
Dull	Intercept	-2.03087	0.197504	105.7336	1	0			
	[sex=1]	-0.84552	0.356421	5.627561	1	0.01768	0.429335	0.213506	0.86334
	[sex=2]	0			0				

2) Multinomická regresia

Všeobecný model

Nech

- závislá premenná Y má r kategórií, a $X_1, \ldots, X_k$ sú vysvetľujúce premenné;

2) Multinomická regresia

Všeobecný model

Nech

- závislá premenná Y má r kategórií, a $X_1, \ldots, X_k$ sú vysvetľujúce premenné;
- $y_i = (y_{i1}, \ldots, y_{ir})$ sú hodnoty závislej premennej v i-tej podskupine, ktoré majú multinomické rozdelenie $Mn(n_i, p_{i1}, \ldots, p_{ir})$;

2) Multinomická regresia

Všeobecný model

Nech

- závislá premenná Y má r kategórií, a $X_1, \ldots, X_k$ sú vysvetľujúce premenné;
- $y_i = (y_{i1}, \ldots, y_{ir})$ sú hodnoty závislej premennej v i-tej podskupine, ktoré majú multinomické rozdelenie $Mn(n_i, p_{i1}, \ldots, p_{ir})$;
- $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{kj})'$ sú regresné koeficienty pre j-tu kategóriu výsledku vzhľadom k j^*-tej (referenčnej) kategórii;

2) Multinomická regresia

Všeobecný model

Nech

- závislá premenná Y má r kategórií, a $X_1, \ldots, X_k$ sú vysvetľujúce premenné;
- $y_i = (y_{i1}, \ldots, y_{ir})$ sú hodnoty závislej premennej v i-tej podskupine, ktoré majú multinomické rozdelenie $Mn(n_i, p_{i1}, \ldots, p_{ir})$;
- $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{kj})'$ sú regresné koeficienty pre j-tu kategóriu výsledku vzhľadom k j^*-tej (referenčnej) kategórii;
- $x_i = (1, x_{i1}, \ldots, x_{ik})'$ sú hodnoty vysvetľujúcich premenných pre i-tu podskupinu.

2) Multinomická regresia

Všeobecný model

Nech

- závislá premenná Y má r kategórií, a $X_1, \ldots, X_k$ sú vysvetľujúce premenné;
- $y_i = (y_{i1}, \ldots, y_{ir})$ sú hodnoty závislej premennej v i-tej podskupine, ktoré majú multinomické rozdelenie $Mn(n_i, p_{i1}, \ldots, p_{ir})$;
- $\beta_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{kj})'$ sú regresné koeficienty pre j-tu kategóriu výsledku vzhľadom k j^*-tej (referenčnej) kategórii;
- $x_i = (1, x_{i1}, \ldots, x_{ik})'$ sú hodnoty vysvetľujúcich premenných pre i-tu podskupinu.

Všeobecný multinomický regresný model má potom tvar

$$\log\left(\frac{p_{ij}}{p_{ij^*}}\right) = x_i'\beta_j\,, \quad j \neq j^*$$

2) Multinomická regresia

- Inverzné vzorce sú

$$p_{ij} = \frac{\exp\left(x_i'\beta_j\right)}{1 + \sum\limits_{\substack{k=1 \\ k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}, \; j \neq j^*$$

a

$$p_{ij^*} = \frac{1}{1 + \sum\limits_{\substack{k=1 \\ k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}.$$

2) Multinomická regresia

- Inverzné vzorce sú

$$p_{ij} = \frac{\exp\left(x_i'\beta_j\right)}{1 + \sum\limits_{\substack{k=1 \\ k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}, \; j \neq j^*$$

a

$$p_{ij^*} = \frac{1}{1 + \sum\limits_{\substack{k=1 \\ k\neq j^*}}^{r} \exp\left(x_i'\beta_k\right)}.$$

- Odhadovanie sa robí opäť pomocou ML-metódy.

2) Multinomická regresia

- Log-vierohodnostná funkcia je

$$\ell(\beta) = \log\left(\frac{n_i!}{\prod_{j=1}^{r} y_{ij!}}\right) + \sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\log\left(p_{ij}\right)$$

2) Multinomická regresia

- Log-vierohodnostná funkcia je

$$\ell(\beta) = \log\left(\frac{n_i!}{\prod_{j=1}^{r} y_{ij!}}\right) + \sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\log\left(p_{ij}\right)$$

- ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta_{mj}} = \sum_{i=1}^{n} x_{im}\left(y_{ij} - n_i p_{ij}\right) = 0, \quad j \neq j^*,\ m = 0, \ldots, k$$

2) Multinomická regresia

- Log-vierohodnostná funkcia je

$$\ell(\beta) = \log\left(\frac{n_i!}{\prod_{j=1}^{r} y_{ij!}}\right) + \sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\log\left(p_{ij}\right)$$

- ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta_{mj}} = \sum_{i=1}^{n} x_{im}\left(y_{ij} - n_i p_{ij}\right) = 0, \quad j \neq j^*,\ m = 0, \ldots, k$$

- Hessián odhadov $\beta = \left(\beta_j', j = 1, \ldots, r, j \neq j^*\right)'$ je

$$H = -\sum_{i=1}^{n}\left(I_{r-1} \otimes x_i\right)\hat{V}_i^*\left(I_{r-1} \otimes x_i\right)',$$

kde $\hat{V}_i^* = n_i\left(\mathrm{diag}\left(\hat{p}_i^*\right) - \hat{p}_i^*\hat{p}_i^{*\prime}\right)$ a $\hat{p}_i^*$ sú vektory odhadov všetkých pravdepodobností p_{ij} okrem p_{ij^*}

2) Multinomická regresia

- Chí-kvadrát odhadovaného modelu je

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(y_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}}$$

Celkový počet neredundantných parametrov modelu je (r-1)(k+1), a preto platí $\chi^2 \approx \chi^2_{(n-k-1)(r-1)}$.

2) Multinomická regresia

- Chí-kvadrát odhadovaného modelu je

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(y_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}}$$

  Celkový počet neredundantných parametrov modelu je (r-1)(k+1), a preto platí $\chi^2 \approx \chi^2_{(n-k-1)(r-1)}$.

- Deviancia odhadovaného modelu je

$$D_m = 2\left(\ell_f - \ell_m\right) = \sum_{i=1}^{n} \sum_{j=1}^{r} y_{ij} \log\left(\frac{y_{ij}}{n_i \hat{p}_{ij}}\right)$$

  Tu opäť platí $D_m \approx \chi^2_{(n-k-1)(r-1)}$.

2) Multinomická regresia

- Chí-kvadrát odhadovaného modelu je

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{(y_{ij} - n_i \hat{p}_{ij})^2}{n_i \hat{p}_{ij}}$$

  Celkový počet neredundantných parametrov modelu je (r-1)(k+1), a preto platí $\chi^2 \approx \chi^2_{(n-k-1)(r-1)}$.

- Deviancia odhadovaného modelu je

$$D_m = 2\left(\ell_f - \ell_m\right) = \sum_{i=1}^{n} \sum_{j=1}^{r} y_{ij} \log\left(\frac{y_{ij}}{n_i \hat{p}_{ij}}\right)$$

  Tu opäť platí $D_m \approx \chi^2_{(n-k-1)(r-1)}$.

- Môžeme používať tie isté pseudo-R^2 štatistiky ako v modeli logistickej regresie.

2) Multinomická regresia

Ak je skutočná variančná matica y_i podstatne väčšia ako $V_i = n_i \left(\text{diag}\left(p_i\right) - p_i p_i'\right)$ (daná multinomickým modelom), hovoríme o **prebytku rozptylu**. V takom prípade môžeme do modelu pridať škálový parameter σ^2, tak že $\text{var}\, y_i = \sigma^2 V_i$.

2) Multinomická regresia

Ak je skutočná variančná matica y_i podstatne väčšia ako $V_i = n_i\left(\text{diag}\left(p_i\right) - p_i p_i'\right)$ (daná multinomickým modelom), hovoríme o **prebytku rozptylu**. V takom prípade môžeme do modelu pridať škálový parameter σ^2, tak že $\text{var}\, y_i = \sigma^2 V_i$.

- Obvyklý (asymptoticky nestranný) odhad σ^2 je

$$\hat{\sigma}^2 = \frac{\chi^2}{(n-k-1)(r-1)}$$

(alebo D_m namiesto χ^2).

2) Multinomická regresia

Ak je skutočná variančná matica y_i podstatne väčšia ako $V_i = n_i\left(\text{diag}\left(p_i\right) - p_i p_i'\right)$ (daná multinomickým modelom), hovoríme o **prebytku rozptylu**. V takom prípade môžeme do modelu pridať škálový parameter σ^2, tak že $\text{var}\, y_i = \sigma^2 V_i$.

- Obvyklý (asymptoticky nestranný) odhad σ^2 je

$$\hat{\sigma}^2 = \frac{\chi^2}{(n-k-1)(r-1)}$$

  (alebo D_m namiesto χ^2).
- Použitie σ^2 nezmení odhady β.

2) Multinomická regresia

Ak je skutočná variančná matica y_i podstatne väčšia ako $V_i = n_i\left(\text{diag}\left(p_i\right) - p_i p_i'\right)$ (daná multinomickým modelom), hovoríme o **prebytku rozptylu**. V takom prípade môžeme do modelu pridať škálový parameter σ^2, tak že $\text{var}\, y_i = \sigma^2 V_i$.

- Obvyklý (asymptoticky nestranný) odhad σ^2 je

$$\hat{\sigma}^2 = \frac{\chi^2}{(n-k-1)(r-1)}$$

  (alebo D_m namiesto χ^2).
- Použitie σ^2 nezmení odhady β.
- Variančná matica odhadov potom je

$$\text{var}\,\hat{\beta} = \hat{\sigma}^2 \left[\sum_{i=1}^{n} \left(I_{r-1} \otimes x_i\right) \hat{V}_i^* \left(I_{r-1} \otimes x_i\right)'\right]^{-1}$$

2) Multinomická regresia

Testy

- Pre ľubovoľnú maticu $L_{q\times k+1}$ plnej hodnosti platí

$$\hat{\beta}_j' L' \left[L \operatorname{var} \hat{\beta}_j L' \right]^{-1} L\hat{\beta}_j \approx \chi^2_q$$

za platnosti $H_0 : L\beta_j = 0$. To dovoľuje testovať jednotlivé regresné koeficienty alebo ich lineárne kombinácie.

2) Multinomická regresia

Testy

- Pre ľubovoľnú maticu $L_{q\times k+1}$ plnej hodnosti platí

$$\hat{\beta}_j' L' \left[L \operatorname{var} \hat{\beta}_j L' \right]^{-1} L\hat{\beta}_j \approx \chi^2_q$$

  za platnosti $H_0 : L\beta_j = 0$. To dovoľuje testovať jednotlivé regresné koeficienty alebo ich lineárne kombinácie.
- LR test vnorených modelov s k_1 a k_2 ($k_1 < k_2$) regresnými parametrami je založený na

$$\frac{1}{\hat{\sigma}^2}\left(D_{m_1} - D_{m_2}\right) = \frac{2}{\hat{\sigma}^2}\left(\hat{\ell}_{m_2} - \hat{\ell}_{m_1}\right) \approx \chi^2_{k_2-k_1}$$

2) Multinomická regresia

Príklad: V štúdii modelujúcej úroveň spokojnosti so životom sme ako prediktor použili pohlavie. Tento prvý model mal 4 neredundantné parametre (konštantu a jednu nereferenčnú kategóriu pohlavia pre každý z dvoch nereferenčných životných pocitov).

2) Multinomická regresia

Príklad: V štúdii modelujúcej úroveň spokojnosti so životom sme ako prediktor použili pohlavie. Tento prvý model mal 4 neredundantné parametre (konštantu a jednu nereferenčnú kategóriu pohlavia pre každý z dvoch nereferenčných životných pocitov).
Ak pridáme rasu (bielu, čiernu, inú) ako ďalší možný prediktor, budeme mať 4 ďalšie neredundantné parametre – 2 nereferenčné rasy pre každý z 2 nereferenčných životných pocitov.

2) Multinomická regresia

Príklad: V štúdii modelujúcej úroveň spokojnosti so životom sme ako prediktor použili pohlavie. Tento prvý model mal 4 neredundantné parametre (konštantu a jednu nereferenčnú kategóriu pohlavia pre každý z dvoch nereferenčných životných pocitov).
Ak pridáme rasu (bielu, čiernu, inú) ako ďalší možný prediktor, budeme mať 4 ďalšie neredundantné parametre – 2 nereferenčné rasy pre každý z 2 nereferenčných životných pocitov.
Prebytok rozptylu nebol pozorovaný. Software dáva:

$$\hat{\ell}_{m_1} = -25.8165, \quad \hat{\ell}_{m_2} = -24.332$$

Z toho

$$\Delta D_m = 2(-24.332 + 25,8165) = 2.969 < 9.488 = \chi^2_{8-4}(0.05)$$

Príslušná p-hodnota je 0.563. Rasový faktor je teda štatisticky nevýznamný.

3) Ordinálna regresia

Príklad: Náhodne vybraná vzorka vermontských občanov mala ohodnotiť prácu kriminalistov v ich štáte. Ponúknutá škála bola Nedostatočná (1), Dostatočná (2), Dobrá (3) a Vynikajúca (4). Súčasťou dotazníka bola aj otázka, či niekto z ich domácnosti bol obeťou kriminálneho činu v priebehu posledných 3 rokov.

3) Ordinálna regresia

Príklad: Náhodne vybraná vzorka vermontských občanov mala ohodnotiť prácu kriminalistov v ich štáte. Ponúknutá škála bola Nedostatočná (1), Dostatočná (2), Dobrá (3) a Vynikajúca (4). Súčasťou dotazníka bola aj otázka, či niekto z ich domácnosti bol obeťou kriminálneho činu v priebehu posledných 3 rokov.

Skúmalo sa, či ľudia s osobnou skúsenosťou so zločinom a ľudia bez tejto skúsenosti hodnotia prácu kriminalistov rovnako.

3) Ordinálna regresia

Príklad: Náhodne vybraná vzorka vermontských občanov mala ohodnotiť prácu kriminalistov v ich štáte. Ponúknutá škála bola Nedostatočná (1), Dostatočná (2), Dobrá (3) a Vynikajúca (4). Súčasťou dotazníka bola aj otázka, či niekto z ich domácnosti bol obeťou kriminálneho činu v priebehu posledných 3 rokov.

Skúmalo sa, či ľudia s osobnou skúsenosťou so zločinom a ľudia bez tejto skúsenosti hodnotia prácu kriminalistov rovnako.

Zistené dáta:

Obeť v domácnosti	Hodnotenie práce kriminalistov				suma
	Nedostatočná	Dostatočná	Dobrá	Vynikajúca	
Áno	14	28	31	3	76
Nie	38	170	248	34	490
suma	52	198	279	37	566

3) Ordinálna regresia

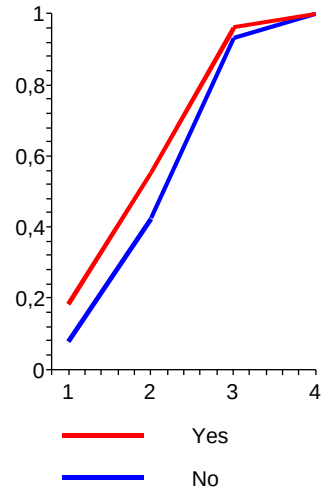
Pri ordinálnych dátach je prirodzené uvažovať pravdepodobnosti kumulatívnych javov, napr. konkrétne hodnotenie alebo horšie. Tabuľka kumulatívnych početností vyzerá takto:

Obeť v domácnosti	Hodnotenie práce kriminalistov			
	Nedostatočná	Dostatočná alebo horšia	Dobrá alebo horšia	Vynikajúca alebo horšia
Áno	14	42	73	76
riadkový podiel	18,42%	55,26%	96,05%	100,00%
Nie	38	208	456	490
riadkový podiel	7,76%	42,45%	93,06%	100,00%

3) Ordinálna regresia

Pri ordinálnych dátach je prirodzené uvažovať pravdepodobnosti kumulatívnych javov, napr. konkrétne hodnotenie alebo horšie. Tabuľka kumulatívnych početností vyzerá takto:

Obeť v domácnosti	Hodnotenie práce kriminalistov			
	Nedostatočná	Dostatočná alebo horšia	Dobrá alebo horšia	Vynikajúca alebo horšia
Áno	14	42	73	76
riadkový podiel	18,42%	55,26%	96,05%	100,00%
Nie	38	208	456	490
riadkový podiel	7,76%	42,45%	93,06%	100,00%

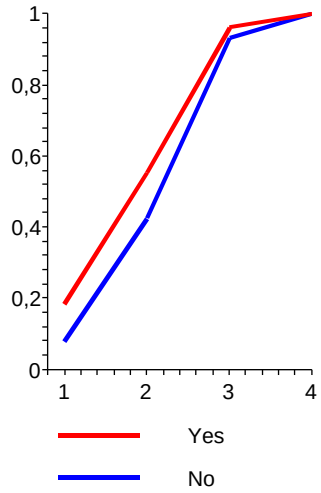


3) Ordinálna regresia

Pri ordinálnych dátach je prirodzené uvažovať pravdepodobnosti kumulatívnych javov, napr. konkrétne hodnotenie alebo horšie. Tabuľka kumulatívnych početností vyzerá takto:

Obeť v domácnosti	Hodnotenie práce kriminalistov			
	Nedostatočná	Dostatočná alebo horšia	Dobrá alebo horšia	Vynikajúca alebo horšia
Áno riadkový podiel	14 18,42%	42 55,26%	73 96,05%	76 100,00%
Nie riadkový podiel	38 7,76%	208 42,45%	456 93,06%	490 100,00%

Graf vytvára dojem, že osobná skúsenosť so zločinom vedie k negatívnejšiemu hodnoteniu práce kriminalistov.
Obe čiary sa prirodzene musia stretnúť na 100%. Ináč vyzerajú prakticky paralelné. Do úvahy teda prichádza model s rovnakým sklonom priamky pre obe kategórie.



3) Ordinálna regresia

Označme $p_{ij}^c = P(\text{hodnotenie} \leq j)$, $i = 1(\text{Nie}), 2(\text{Áno})$, $j = 1, 2, 3$ netriviálne kumulatívne pravdepodobnosti. Náš model potom je

$$\log\left(\frac{p_{1j}^c}{1-p_{1j}^c}\right) = \alpha_j \quad \text{a} \quad \log\left(\frac{p_{2j}^c}{1-p_{2j}^c}\right) = \alpha_j + \beta\,,$$

3) Ordinálna regresia

Označme $p_{ij}^c = P(\text{hodnotenie} \leq j)$, $i = 1(\text{Nie}), 2(\text{Áno})$, $j = 1, 2, 3$ netriviálne kumulatívne pravdepodobnosti. Náš model potom je

$$\log\left(\frac{p_{1j}^c}{1-p_{1j}^c}\right) = \alpha_j \quad \text{a} \quad \log\left(\frac{p_{2j}^c}{1-p_{2j}^c}\right) = \alpha_j + \beta\,,$$

alebo

$$\log\left(\frac{p_j^c(x)}{1-p_j^c(x)}\right) = \alpha_j + \beta x \quad \forall j\,, x \in \{0, 1\}.$$

3) Ordinálna regresia

Označme $p_{ij}^c = P(\text{hodnotenie} \leq j)$, $i = 1(\text{Nie}), 2(\text{Áno})$, $j = 1, 2, 3$ netriviálne kumulatívne pravdepodobnosti. Náš model potom je

$$\log\left(\frac{p_{1j}^c}{1-p_{1j}^c}\right) = \alpha_j \quad \text{a} \quad \log\left(\frac{p_{2j}^c}{1-p_{2j}^c}\right) = \alpha_j + \beta\,,$$

alebo

$$\log\left(\frac{p_j^c(x)}{1-p_j^c(x)}\right) = \alpha_j + \beta x \quad \forall j\,, x \in \{0,1\}.$$

Software dáva $\alpha_1 = -2.39$, $\alpha_2 = -0.32$, $\alpha_3 = 2.59$, $\beta = 0.63$. Pomocou štandardného vzorca pre inverziu logitov dostaneme nasledujúce odhady:

	≤ 1	≤ 2	≤ 3
Áno	14,69%	57,85%	96,18%
Nie	8,38%	42,15%	93,04%

3) Ordinálna regresia

Model proporcionálnych šancí

Nech

- Y je ordinálna závislá premenná s možnými hodnotami $1, \ldots, r$

3) Ordinálna regresia

Model proporcionálnych šancí

Nech

- Y je ordinálna závislá premenná s možnými hodnotami $1, \ldots, r$
- $X = (X_1, \ldots, X_k)$ sú nezávislé regresné premenné

3) Ordinálna regresia

Model proporcionálnych šancí

Nech

- Y je ordinálna závislá premenná s možnými hodnotami $1, \ldots, r$
- $X = (X_1, \ldots, X_k)$ sú nezávislé regresné premenné
- $\alpha_1, \ldots, \alpha_{r-1}$ a $\beta = (\beta_1, \ldots, \beta_k)$ sú neznáme regresné koeficienty

3) Ordinálna regresia

Model proporcionálnych šancí

Nech

- Y je ordinálna závislá premenná s možnými hodnotami $1, \ldots, r$
- $X = (X_1, \ldots, X_k)$ sú nezávislé regresné premenné
- $\alpha_1, \ldots, \alpha_{r-1}$ a $\beta = (\beta_1, \ldots, \beta_k)$ sú neznáme regresné koeficienty

Model: logity kumulatívnych pravdepodobností
$p_j^c(x) = P(Y \leq j | X = x)$ spĺňajú

$$\log\left(\frac{p_j^c(x)}{1 - p_j^c(x)}\right) = \alpha_j + \beta' x \quad \forall j = 1, \ldots, r-1$$

3) Ordinálna regresia

Model proporcionálnych šancí

Nech

- Y je ordinálna závislá premenná s možnými hodnotami $1, \ldots, r$
- $X = (X_1, \ldots, X_k)$ sú nezávislé regresné premenné
- $\alpha_1, \ldots, \alpha_{r-1}$ a $\beta = (\beta_1, \ldots, \beta_k)$ sú neznáme regresné koeficienty

Model: logity kumulatívnych pravdepodobností
$p_j^c(x) = P(Y \leq j | X = x)$ spĺňajú

$$\log\left(\frac{p_j^c(x)}{1 - p_j^c(x)}\right) = \alpha_j + \beta' x \quad \forall j = 1, \ldots, r-1$$

Pretože logaritmus kumulatívneho pomeru šancí dosiahnutia rovnakej hodnoty závislej premennej v rôznych x-ových bodoch je priamo úmerný ich vzdialenosti, model nazývame modelom proporcionálnych šancí:

$$\log\left(\frac{p_j^c(x_1)}{1 - p_j^c(x_1)} \cdot \frac{1 - p_j^c(x_2)}{p_j^c(x_2)}\right) = \beta'(x_1 - x_2)$$

3) Ordinálna regresia

- Odhadovanie parametrov sa robí opäť ML-metódou. Sú pritom dve možnosti:

3) Ordinálna regresia

- Odhadovanie parametrov sa robí opäť ML-metódou. Sú pritom dve možnosti:
  1. najprv odhadnúť $p_j^c(x)$, a potom vyrátať $p_j(x) = p_j^c(x) - p_{j-1}^c(x)$ (definujeme $p_0(x) \equiv 0$)

3) Ordinálna regresia

- Odhadovanie parametrov sa robí opäť ML-metódou. Sú pritom dve možnosti:
  1. najprv odhadnúť $p_j^c(x)$, a potom vyrátať $p_j(x) = p_j^c(x) - p_{j-1}^c(x)$ (definujeme $p_0(x) \equiv 0$)
  2. odhadovať priamo $p_j(x)$

3) Ordinálna regresia

- Odhadovanie parametrov sa robí opäť ML-metódou. Sú pritom dve možnosti:
  1. najprv odhadnúť $p_j^c(x)$, a potom vyrátať $p_j(x) = p_j^c(x) - p_{j-1}^c(x)$ (definujeme $p_0(x) \equiv 0$)
  2. odhadovať priamo $p_j(x)$
- Vierohodnostná funkcia je

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[p_j\left(x_i\right) - p_{j-1}\left(x_i\right)\right]^{y_{ij}}$$

3) Ordinálna regresia

- Odhadovanie parametrov sa robí opäť ML-metódou. Sú pritom dve možnosti:
  1. najprv odhadnúť $p_j^c(x)$, a potom vyrátať $p_j(x) = p_j^c(x) - p_{j-1}^c(x)$ (definujeme $p_0(x) \equiv 0$)
  2. odhadovať priamo $p_j(x)$
- Vierohodnostná funkcia je

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[p_j\left(x_i\right) - p_{j-1}\left(x_i\right)\right]^{y_{ij}}$$

- Všetky štandardné miery dobrej zhody sú použiteľné.

3) Ordinálna regresia

- Odhadovanie parametrov sa robí opäť ML-metódou. Sú pritom dve možnosti:
  1. najprv odhadnúť $p_j^c(x)$, a potom vyrátať $p_j(x) = p_j^c(x) - p_{j-1}^c(x)$ (definujeme $p_0(x) \equiv 0$)
  2. odhadovať priamo $p_j(x)$
- Vierohodnostná funkcia je

$$L(\alpha, \beta) = \prod_{i=1}^{n} \prod_{j=1}^{r} \left[p_j\left(x_i\right) - p_{j-1}\left(x_i\right)\right]^{y_{ij}}$$

- Všetky štandardné miery dobrej zhody sú použiteľné.
- Odhady parametrov sú iné ako pri separátnych logitových modeloch pre všetky j.

3) Ordinálna regresia

Example: Výstup software pre vermontské kriminalistické dáta:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,39221	0,15177	248,44332	1	0,00000	-2,68968	-2,09475
	[rating = 2]	-0,31651	0,09082	12,14637	1	0,00049	-0,49451	-0,13852
	[rating = 3]	2,59316	0,17163	228,28667	1	0,00000	2,25678	2,92955
Location	[hhcrime=1]	-0,63298	0,23198	7,44539	1	0,00636	-1,08765	-0,17831
	[hhcrime=2]	0	.	.	0	.	.	.

Všimnite si opačných znamienok koeficienta β (hhcrime=1). Veľa softwarov používa model $\alpha_j - \beta x$ kvôli lepšej interpretácii: v takomto modeli totiž väčšie koeficienty sú asociované s vyšším hodnotením.

3) Ordinálna regresia

Example: Výstup software pre vermontské kriminalistické dáta:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,39221	0,15177	248,44332	1	0,00000	-2,68968	-2,09475
	[rating = 2]	-0,31651	0,09082	12,14637	1	0,00049	-0,49451	-0,13852
	[rating = 3]	2,59316	0,17163	228,28667	1	0,00000	2,25678	2,92955
Location	[hhcrime=1]	-0,63298	0,23198	7,44539	1	0,00636	-1,08765	-0,17831
	[hhcrime=2]	0	.	.	0	.	.	.

Všimnite si opačných znamienok koeficienta β (hhcrime=1). Veľa softwarov používa model $\alpha_j - \beta x$ kvôli lepšej interpretácii: v takomto modeli totiž väčšie koeficienty sú asociované s vyšším hodnotením.

Pridajme teraz ďalší prediktor, pohlavie:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,57419	0,17641	212,93519	1	0,00000	-2,91995	-2,22844
	[rating = 2]	-0,48730	0,12326	15,62868	1	0,00008	-0,72890	-0,24571
	[rating = 3]	2,43740	0,18672	170,40298	1	0,00000	2,07143	2,80336
Location	[hhcrime=1]	-0,62074	0,23228	7,14177	1	0,00753	-1,07599	-0,16548
	[hhcrime=2]	0	.	.	0	.	.	.
	[sex=1]	-0,34145	0,16030	4,53709	1	0,03317	-0,65563	-0,02726
	[sex=2]	0	.	.	0	.	.	.

3) Ordinálna regresia

Máme podozrenie, že pohlavie môže ovplyvniť senzitivitu k obetiam zločinu, takže pridáme aj interakciu:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,64904	0,18097	214,26179	1	0,00000	-3,00374	-2,29434
	[rating = 2]	-0,55150	0,12873	18,35418	1	0,00002	-0,80381	-0,29920
	[rating = 3]	2,38107	0,18819	160,07877	1	0,00000	2,01222	2,74993
Location	[hhcrime=1]	-1,13654	0,33008	11,85565	1	0,00057	-1,78350	-0,48959
	[hhcrime=2]	0	.	.	0	.	.	.
	[sex=1]	-0,46925	0,17330	7,33183	1	0,00677	-0,80891	-0,12959
	[sex=2]	0	.	.	0	.	.	.
	[hhcrime=1] * [sex=1]	0,95889	0,46413	4,26832	1	0,03883	0,04921	1,86857
	[hhcrime=1] * [sex=2]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=1]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=2]	0	.	.	0	.	.	.

3) Ordinálna regresia

Máme podozrenie, že pohlavie môže ovplyvniť senzitivitu k obetiam zločinu, takže pridáme aj interakciu:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,64904	0,18097	214,26179	1	0,00000	-3,00374	-2,29434
	[rating = 2]	-0,55150	0,12873	18,35418	1	0,00002	-0,80381	-0,29920
	[rating = 3]	2,38107	0,18819	160,07877	1	0,00000	2,01222	2,74993
Location	[hhcrime=1]	-1,13654	0,33008	11,85565	1	0,00057	-1,78350	-0,48959
	[hhcrime=2]	0	.	.	0	.	.	.
	[sex=1]	-0,46925	0,17330	7,33183	1	0,00677	-0,80891	-0,12959
	[sex=2]	0	.	.	0	.	.	.
	[hhcrime=1] * [sex=1]	0,95889	0,46413	4,26832	1	0,03883	0,04921	1,86857
	[hhcrime=1] * [sex=2]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=1]	0	.	.	0	.	.	.
	[hhcrime=2] * [sex=2]	0	.	.	0	.	.	.

Keďže interakcia je štatisticky významná, jej individuálne zložky nemajú samostatne dobrý zmysel:

		Estimate	Std. Error	Wald	df	Sig.	95% conf. interval	
Threshold	[rating = 1]	-2,64904	0,18097	214,26179	1	0,00000	-3,00374	-2,29434
	[rating = 2]	-0,55150	0,12873	18,35418	1	0,00002	-0,80381	-0,29920
	[rating = 3]	2,38107	0,18819	160,07877	1	0,00000	2,01222	2,74993
Location	[hhcrime=1] * [sex=1]	-0,64690	0,32950	3,85460	1	0,04961	-1,29270	-0,00110
	[hhcrime=1] * [sex=2]	-1,13654	0,33008	11,85565	1	0,00057	-1,78350	-0,48959
	[hhcrime=2] * [sex=1]	-0,46925	0,17330	7,33183	1	0,00677	-0,80891	-0,12959
	[hhcrime=2] * [sex=2]	0	.	.	0	.	.	.

Redundantné parametre sa neodhadujú, stačí samotná interakcia. Tento model má to isté χ^2, devianciu a pseudo-R^2 ako predchádzajúci.

3) Ordinálna regresia

Ďalšie modely ordinálnej regresie

- Všeobecný kumulatívny logitový model je

$$\log\left(\frac{p_j^c(x)}{1-p_j^c(x)}\right) = \alpha_j + \beta_j' x \quad \forall j = 1, \ldots, r-1$$

V tomto modeli má každá skupina svoju vlastnú smernicu. Model proporcionálnych šancí je iba špeciálnym prípadom, ktorý sa dá testovať LR-testom.

3) Ordinálna regresia

Ďalšie modely ordinálnej regresie

- Všeobecný kumulatívny logitový model je

$$\log\left(\frac{p_j^c(x)}{1-p_j^c(x)}\right) = \alpha_j + \beta_j' x \quad \forall j = 1, \ldots, r-1$$

  V tomto modeli má každá skupina svoju vlastnú smernicu. Model proporcionálnych šancí je iba špeciálnym prípadom, ktorý sa dá testovať LR-testom.
- Model susedných kategórií je

$$\log\left(\frac{p_j(x)}{p_{j+1}(x)}\right) = \alpha_j + \beta' x \quad \forall j = 1, \ldots, r-1$$

  Tento model tiež využíva usporiadanie, lebo

$$\log\left(\frac{p_j(x)}{p_r(x)}\right) = \sum_{m=j}^{r} \alpha_m + \beta'(r-j)x \quad \forall j$$

4) Poissonova regresia

Príklad: Doll (1971) skúmal vzťah medzi fajčením a rakovinou pľúc u britských lekárov.

4) Poissonova regresia

Príklad: Doll (1971) skúmal vzťah medzi fajčením a rakovinou pľúc u britských lekárov.

Zisťoval, do akej miery môže intenzita fajčenia (ak vôbec) vysvetliť pozorované množstvo prípadov rakoviny pľúc.

4) Poissonova regresia

Príklad: Doll (1971) skúmal vzťah medzi fajčením a rakovinou pľúc u britských lekárov.

Zisťoval, do akej miery môže intenzita fajčenia (ak vôbec) vysvetliť pozorované množstvo prípadov rakoviny pľúc.

Dáta pre podskupinu fajčiarov 45–49 ročných):

počet cigariet denne (v priemere)	0	5.2	11.2	15.9	20.4	27.4	40.8
počet prípadov	0	0	2	2	9	10	7
počet osobo-rokov	1421	927	988	849	1567	1409	556

4) Poissonova regresia

Príklad: Doll (1971) skúmal vzťah medzi fajčením a rakovinou pľúc u britských lekárov.

Zisťoval, do akej miery môže intenzita fajčenia (ak vôbec) vysvetliť pozorované množstvo prípadov rakoviny pľúc.

Dáta pre podskupinu fajčiarov 45–49 ročných):

počet cigariet denne (v priemere)	0	5.2	11.2	15.9	20.4	27.4	40.8
počet prípadov	0	0	2	2	9	10	7
počet osobo-rokov	1421	927	988	849	1567	1409	556

- Sledovaná premenná nadobúda nezáporné celočíselné hodnoty.

4) Poissonova regresia

Príklad: Doll (1971) skúmal vzťah medzi fajčením a rakovinou pľúc u britských lekárov.

Zisťoval, do akej miery môže intenzita fajčenia (ak vôbec) vysvetliť pozorované množstvo prípadov rakoviny pľúc.

Dáta pre podskupinu fajčiarov 45–49 ročných):

počet cigariet denne (v priemere)	0	5.2	11.2	15.9	20.4	27.4	40.8
počet prípadov	0	0	2	2	9	10	7
počet osobo-rokov	1421	927	988	849	1567	1409	556

- Sledovaná premenná nadobúda nezáporné celočíselné hodnoty.
- Početnosti sú generované Poissonovým procesom.

4) Poissonova regresia

Príklad: Doll (1971) skúmal vzťah medzi fajčením a rakovinou pľúc u britských lekárov.

Zisťoval, do akej miery môže intenzita fajčenia (ak vôbec) vysvetliť pozorované množstvo prípadov rakoviny pľúc.

Dáta pre podskupinu fajčiarov 45–49 ročných):

počet cigariet denne (v priemere)	0	5.2	11.2	15.9	20.4	27.4	40.8
počet prípadov	0	0	2	2	9	10	7
počet osobo-rokov	1421	927	988	849	1567	1409	556

- Sledovaná premenná nadobúda nezáporné celočíselné hodnoty.
- Početnosti sú generované Poissonovým procesom.
- Každá pozorovaná početnosť závisí na počte pozorovaných osôb (v príslušnej kategórii) a čase.

4) Poissonova regresia

Základné predpoklady modelu sú:

- početnosti generuje Poissonov proces,

4) Poissonova regresia

Základné predpoklady modelu sú:

- početnosti generuje Poissonov proces,
  - počty prípadov v rôznych časových intervaloch sú nazávislé n.v.

4) Poissonova regresia

Základné predpoklady modelu sú:

- početnosti generuje Poissonov proces,
  - počty prípadov v rôznych časových intervaloch sú nazávislé n.v.
  - pravdepodobnosť jednej udalosti v krátkom časovom intervale $\langle t; t+h)$ je $\lambda h + o(h)$, $h \to 0$, $\lambda > 0$ (intenzita udalostí)

4) Poissonova regresia

Základné predpoklady modelu sú:

- početnosti generuje Poissonov proces,
  - počty prípadov v rôznych časových intervaloch sú nazávislé n.v.
  - pravdepodobnosť jednej udalosti v krátkom časovom intervale $\langle t; t+h)$ je $\lambda h + o(h)$, $h \to 0$, $\lambda > 0$ (intenzita udalostí)
  - pravdepodobnosť dvoch alebo viacerých udalostí v krátkom časovom intervale $\langle t; t+h)$ je $o(h)$, $h \to 0$

4) Poissonova regresia

Základné predpoklady modelu sú:

- početnosti generuje Poissonov proces,
  - počty prípadov v rôznych časových intervaloch sú nazávislé n.v.
  - pravdepodobnosť jednej udalosti v krátkom časovom intervale $\langle t; t+h)$ je $\lambda h + o(h)$, $h \to 0$, $\lambda > 0$ (intenzita udalostí)
  - pravdepodobnosť dvoch alebo viacerých udalostí v krátkom časovom intervale $\langle t; t+h)$ je $o(h)$, $h \to 0$
- intenzita udalostí je a (nezáporná) funkcia vysvetľujúcej premennej,

4) Poissonova regresia

Základné predpoklady modelu sú:

- početnosti generuje Poissonov proces,
  - počty prípadov v rôznych časových intervaloch sú nazávislé n.v.
  - pravdepodobnosť jednej udalosti v krátkom časovom intervale $\langle t; t+h)$ je $\lambda h + o(h)$, $h \to 0$, $\lambda > 0$ (intenzita udalostí)
  - pravdepodobnosť dvoch alebo viacerých udalostí v krátkom časovom intervale $\langle t; t+h)$ je $o(h)$, $h \to 0$
- intenzita udalostí je a (nezáporná) funkcia vysvetľujúcej premennej,
- intenzita udalostí sa riadi *the multiplikatívnym modelom*

$$\lambda(X) = \exp\{\beta_0 + \beta_1 X\}$$

4) Poissonova regresia

Features:

- ML-odhadom intenzity udalostí u Poissonovho procesu je počet udalostí delený celkovým pozorovaným osobo-časom, i.e. incidenciou udalostí;

4) Poissonova regresia

Features:

- ML-odhadom intenzity udalostí u Poissonovho procesu je počet udalostí delený celkovým pozorovaným osobo-časom, i.e. incidenciou udalostí;
- uvažujeme podmienené Poissonove procesy pri daných hodnotách vysvetľujúcej premennej;

4) Poissonova regresia

Features:

- ML-odhadom intenzity udalostí u Poissonovho procesu je počet udalostí delený celkovým pozorovaným osobo-časom, i.e. incidenciou udalostí;
- uvažujeme podmienené Poissonove procesy pri daných hodnotách vysvetľujúcej premennej;
- multiplikatívny model dáva vždy kladné odhady λ

4) Poissonova regresia

Features:

- ML-odhadom intenzity udalostí u Poissonovho procesu je počet udalostí delený celkovým pozorovaným osobo-časom, i.e. incidenciou udalostí;
- uvažujeme podmienené Poissonove procesy pri daných hodnotách vysvetľujúcej premennej;
- multiplikatívny model dáva vždy kladné odhady λ
- keďže strednou hodnotou Poissonovho procesu $Po(\lambda t)$ je λt, ľahko sa odvodí

$$\beta_0 = \log E(Y|X = 0) - \log t_0$$

$$\beta_1 = \log \frac{E(Y|X = k+1)}{E(Y|X = k)} - \log \frac{t_{k+1}}{t_k}$$

4) Poissonova regresia

Features:

- ML-odhadom intenzity udalostí u Poissonovho procesu je počet udalostí delený celkovým pozorovaným osobo-časom, i.e. incidenciou udalostí;
- uvažujeme podmienené Poissonove procesy pri daných hodnotách vysvetľujúcej premennej;
- multiplikatívny model dáva vždy kladné odhady λ
- keďže strednou hodnotou Poissonovho procesu $Po(\lambda t)$ je λt, ľahko sa odvodí

$$\beta_0 = \log E(Y|X=0) - \log t_0$$

$$\beta_1 = \log \frac{E(Y|X=k+1)}{E(Y|X=k)} - \log \frac{t_{k+1}}{t_k}$$

Podobnosť s logistickou regresiou: β_0 je log(základná incidencia), a β_1 je log(pomer incidencií) pri jednotkovej zmene vysvetľujúcej premennej, až na multiplikatívnu konštantu.

4) Poissonova regresia

- Model sa obvykle píše v tvare

$$\log E(Y|X) - \log t(X) = \beta_0 + \beta_1 X\,.$$

Množstvo osobo-času $t(X)$ pri hodnote X (doba vystavenia) sa nazýva *offset*.

4) Poissonova regresia

- Model sa obvykle píše v tvare

$$\log E(Y|X) - \log t(X) = \beta_0 + \beta_1 X\,.$$

  Množstvo osobo-času $t(X)$ pri hodnote X (doba vystavenia) sa nazýva *offset*.

- Parametre sa odhadujú numericky pomocou ML-metódy. Log-vierohodnosť modelu je

$$\ell(\beta_0, \beta_1) = -\sum_{i=1}^{n} t\,(X_i) \exp\{\beta_0 + \beta_1 X_i\} + \sum_{i=1}^{n} Y_i\,(\beta_0 + \beta_1 X_i)$$
$$+ \sum_{i=1}^{n} Y_i \log t\,(X_i) - \sum_{i=1}^{n} \log\,(Y_i!)$$

4) Poissonova regresia

- ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left[Y_i - t\left(X_i\right) \exp\left\{\beta_0 + \beta_1 X_i\right\} \right] = 0$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left[Y_i - t\left(X_i\right) \exp\left\{\beta_0 + \beta_1 X_i\right\} \right] X_i = 0$$

4) Poissonova regresia

- ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left[Y_i - t\left(X_i\right) \exp\left\{\beta_0 + \beta_1 X_i\right\} \right] = 0$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left[Y_i - t\left(X_i\right) \exp\left\{\beta_0 + \beta_1 X_i\right\} \right] X_i = 0$$

Software dáva ($R^2_{MF} = 0.4986$)

	coefficient	*std. error*	*z*	*P* > |*z*|	*95% conf. interval*	
cigarettes per day	*0.074846*	*0.015639*	*4.79*	*0.000*	*0.044193*	*0.10550*
constant	*-7.139276*	*0.454019*	*-15.72*	*0.000*	*-8.02914*	*-6.24942*

4) Poissonova regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná generovaná (podmieneným) Poissonovým procesom,

4) Poissonova regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná generovaná (podmieneným) Poissonovým procesom,
- $X = (X_1, \ldots, X_k)$ sú nezávislé vysvetľujúce premenné, ktoré ovplyvňujú intenzitu Poissonovho procesu:

$$\lambda(X) = \exp\{\beta' X\},$$

4) Poissonova regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná generovaná (podmieneným) Poissonovým procesom,
- $X = (X_1, \ldots, X_k)$ sú nezávislé vysvetľujúce premenné, ktoré ovplyvňujú intenzitu Poissonovho procesu:

$$\lambda(X) = \exp\{\beta' X\},$$

- $\beta = (\beta_1, \ldots, \beta_k)$ sú neznáme regresné koeficienty.

4) Poissonova regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná generovaná (podmieneným) Poissonovým procesom,
- $X = (X_1, \ldots, X_k)$ sú nezávislé vysvetľujúce premenné, ktoré ovplyvňujú intenzitu Poissonovho procesu:

$$\lambda(X) = \exp\{\beta' X\},$$

- $\beta = (\beta_1, \ldots, \beta_k)$ sú neznáme regresné koeficienty.

Všeobecný Poissonov regresný model má potom tvar

$$\log E(Y|X) - \log t(X) = \beta' X.$$

4) Poissonova regresia

- Log-vierohodnosť modelu je

$$\ell(\beta) = -\sum_{i=1}^{n} t(X_i) \exp\{\beta' X_i\} + \sum_{i=1}^{n} Y_i (\beta' X_i) + \sum_{i=1}^{n} Y_i \log t(X_i) - \sum_{i=1}^{n} \log(Y_i!)$$

4) Poissonova regresia

- Log-vierohodnosť modelu je

$$\ell(\beta) = -\sum_{i=1}^{n} t(X_i) \exp\{\beta' X_i\} + \sum_{i=1}^{n} Y_i (\beta' X_i) + \sum_{i=1}^{n} Y_i \log t(X_i) - \sum_{i=1}^{n} \log(Y_i!)$$

- ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} [Y_i - t(X_i) \exp\{\beta' X_i\}] X_i = 0$$

4) Poissonova regresia

- Log-vierohodnosť modelu je

$$\ell(\beta) = -\sum_{i=1}^{n} t(X_i) \exp\{\beta' X_i\} + \sum_{i=1}^{n} Y_i (\beta' X_i) + \sum_{i=1}^{n} Y_i \log t(X_i) - \sum_{i=1}^{n} \log(Y_i!)$$

- ML-rovnice sú

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} [Y_i - t(X_i) \exp\{\beta' X_i\}] X_i = 0$$

- Asymptotická variančná matica je

$$\operatorname{var} \hat{\beta} = \left(\sum_{i=1}^{n} t(X_i) \exp\{\beta' X_i\} X_i X_i' \right)^{-1}$$

4) Poissonova regresia

- Použiteľné sú všetky štandardné miery dobrej zhody:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{\hat{Y}_i}, \quad D_m = 2\sum_{i=1}^{n} Y_i \log\left(\frac{Y_i}{\hat{Y}_i}\right) - \left(Y_i - \hat{Y}_i\right).$$

Obidve majú asymptotické rozdelenie $\chi^2(n-k-1)$.

4) Poissonova regresia

- Použiteľné sú všetky štandardné miery dobrej zhody:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{\hat{Y}_i}, \quad D_m = 2\sum_{i=1}^{n} Y_i \log\left(\frac{Y_i}{\hat{Y}_i}\right) - \left(Y_i - \hat{Y}_i\right).$$

  Obidve majú asymptotické rozdelenie $\chi^2(n-k-1)$.

Príklad: Uvažujme kompletné údaje Dolla (1971) o fajčení a rakovine pľúc. Potom môžeme vziať dva prediktory – priemerný počet cigariet denne a priemerný počet rokov fajčenia v príslušnej skupine.

4) Poissonova regresia

- Použiteľné sú všetky štandardné miery dobrej zhody:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{\hat{Y}_i}, \quad D_m = 2\sum_{i=1}^{n} Y_i \log\left(\frac{Y_i}{\hat{Y}_i}\right) - \left(Y_i - \hat{Y}_i\right).$$

  Obidve majú asymptotické rozdelenie $\chi^2(n-k-1)$.

Príklad: Uvažujme kompletné údaje Dolla (1971) o fajčení a rakovine pľúc. Potom môžeme vziať dva prediktory – priemerný počet cigariet denne a priemerný počet rokov fajčenia v príslušnej skupine.

Software dáva ($R^2_{MF} = 0.6092$)

	coefficient	*std. error*	*z*	$P > \|z\|$	*95% conf. interval*	
cigarettes per day	*0.0789375*	*0.0058775*	*13.43*	*0.000*	*0.0674177*	*0.0904573*
years of smoking	*0.1064206*	*0.0076236*	*13.96*	*0.000*	*0.0914786*	*0.1213626*
constant	*-12.31893*	*0.3859998*	*-31.91*	*0.000*	*-13.07548*	*-11.56239*

4) Poissonova regresia

Prebytok rozptylu

Podobne ako v predchádzajúcom modeli, aj v tomto môžeme pozorovať prebytok (príp. nedostatok) rozptylu. Najjednoduchším riešením je opäť zavedenie škálového parametra σ^2.

4) Poissonova regresia

Prebytok rozptylu

Podobne ako v predchádzajúcom modeli, aj v tomto môžeme pozorovať prebytok (príp. nedostatok) rozptylu. Najjednoduchším riešením je opäť zavedenie škálového parametra σ^2.
Možné odhady sú

$$\hat{\sigma}^2 = \frac{1}{n-k}\chi^2 \quad \text{a} \quad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\hat{Y}_i}.$$

ML-odhad $\hat{\beta}$ nezávisí na σ^2 a jeho variancia je σ^2-násobok variancie štandardného modelu.

4) Poissonova regresia

Prebytok rozptylu

Podobne ako v predchádzajúcom modeli, aj v tomto môžeme pozorovať prebytok (príp. nedostatok) rozptylu. Najjednoduchším riešením je opäť zavedenie škálového parametra σ^2.
Možné odhady sú

$$\hat{\sigma}^2 = \frac{1}{n-k}\chi^2 \quad \text{a} \quad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\hat{Y}_i}.$$

ML-odhad $\hat{\beta}$ nezávisí na σ^2 a jeho variancia je σ^2-násobok variancie štandardného modelu.

Inou možnosťou je použiť **negatívne binomickú regresiu**.

4) Poissonova regresia

Prebytok rozptylu

Podobne ako v predchádzajúcom modeli, aj v tomto môžeme pozorovať prebytok (príp. nedostatok) rozptylu. Najjednoduchším riešením je opäť zavedenie škálového parametra σ^2.
Možné odhady sú

$$\hat{\sigma}^2 = \frac{1}{n-k}\chi^2 \quad \text{a} \quad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\hat{Y}_i}.$$

ML-odhad $\hat{\beta}$ nezávisí na σ^2 a jeho variancia je σ^2-násobok variancie štandardného modelu.

Inou možnosťou je použiť **negatívne binomickú regresiu**.

Modely Poissonovej regresie **bez núl** a **s prebytkom núl** riešia problémy s chýbajúcimi alebo príliš mnohými nulami.

5) Negatívne binomická regresia

Vzťah k Poissonovej regresii

Model NB1:

- Nech Y má Poissonovo rozdelenie s náhodnou strednou hodnotou Z, kde $Z \sim \Gamma(\alpha\mu, 1/\alpha)$, t.j. hustota Z je

$$f(x) = \frac{\alpha^{\alpha\mu}}{\Gamma(\alpha\mu)} x^{\alpha\mu-1} e^{-\alpha x}, \quad x > 0.$$

5) Negatívne binomická regresia

Vzťah k Poissonovej regresii

Model NB1:

- Nech Y má Poissonovo rozdelenie s náhodnou strednou hodnotou Z, kde $Z \sim \Gamma(\alpha\mu, 1/\alpha)$, t.j. hustota Z je

$$f(x) = \frac{\alpha^{\alpha\mu}}{\Gamma(\alpha\mu)} x^{\alpha\mu-1} e^{-\alpha x}, \quad x > 0.$$

V takom prípade $E\,Z = \mu$, $\mathrm{var}\,Z = \mu/\alpha$. Platí teda

$$P(Y = y) = \frac{\Gamma(y + \alpha\mu)}{y!\Gamma(\alpha\mu)} \left(\frac{\alpha}{1+\alpha}\right)^{\alpha\mu} \left(\frac{1}{1+\alpha}\right)^{y},$$

čiže Y má negatívne binomické rozdelenie $NB(\alpha/(1+\alpha), \alpha\mu)$.

5) Negatívne binomická regresia

Vzťah k Poissonovej regresii

Model NB1:

- Nech Y má Poissonovo rozdelenie s náhodnou strednou hodnotou Z, kde $Z \sim \Gamma(\alpha\mu, 1/\alpha)$, t.j. hustota Z je

$$f(x) = \frac{\alpha^{\alpha\mu}}{\Gamma(\alpha\mu)} x^{\alpha\mu-1} e^{-\alpha x}, \quad x > 0.$$

V takom prípade $E\,Z = \mu$, $\mathrm{var}\,Z = \mu/\alpha$. Platí teda

$$P(Y = y) = \frac{\Gamma(y + \alpha\mu)}{y!\Gamma(\alpha\mu)} \left(\frac{\alpha}{1+\alpha}\right)^{\alpha\mu} \left(\frac{1}{1+\alpha}\right)^{y},$$

čiže Y má negatívne binomické rozdelenie $NB(\alpha/(1+\alpha), \alpha\mu)$. Z toho vyplýva

$$E\,Y = \mu \quad \text{a} \quad \mathrm{var}\,Y = \mu\left(1 + \tfrac{1}{\alpha}\right).$$

Rozptyl sa teda zväčší $\left(1 + \frac{1}{\alpha}\right)$-krát.

5) Negatívne binomická regresia

Model NB2:

- Nech Y má Poissonovo rozdelenie s náhodnou strednou hodnotou μU, kde $U \sim \Gamma(\alpha, 1/\alpha)$, t.j. hustota U je

$$f(x) = \frac{\alpha^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x}, \quad x > 0.$$

5) Negatívne binomická regresia

Model NB2:

- Nech Y má Poissonovo rozdelenie s náhodnou strednou hodnotou μU, kde $U \sim \Gamma(\alpha, 1/\alpha)$, t.j. hustota U je

$$f(x) = \frac{\alpha^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x}, \quad x > 0.$$

V tom prípade $E\,U = 1$, $\text{var}\,U = 1/\alpha$. platí teda

$$P(Y = y) = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \left(\frac{\alpha}{\mu+\alpha}\right)^\alpha \left(\frac{\mu}{\mu+\alpha}\right)^y,$$

čiže Y má negatívne binomické rozdelenie $NB(\alpha/(\mu+\alpha), \alpha)$.

5) Negatívne binomická regresia

Model NB2:

- Nech Y má Poissonovo rozdelenie s náhodnou strednou hodnotou μU, kde $U \sim \Gamma(\alpha, 1/\alpha)$, t.j. hustota U je

$$f(x) = \frac{\alpha^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x}, \quad x > 0.$$

V tom prípade $E\,U = 1$, $\text{var}\,U = 1/\alpha$. platí teda

$$P(Y = y) = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \left(\frac{\alpha}{\mu+\alpha}\right)^\alpha \left(\frac{\mu}{\mu+\alpha}\right)^y,$$

čiže Y má negatívne binomické rozdelenie $NB(\alpha/(\mu+\alpha), \alpha)$. Z toho

$$E\,Y = \mu \quad \text{a} \quad \text{var}\,Y = \mu\left(1 + \tfrac{\mu}{\alpha}\right).$$

Rozptyl sa teda zväčší $\left(1 + \frac{\mu}{\alpha}\right)$-krát a faktor prebytku rozptylu sa mení podľa strednej hodnoty.

5) Negatívne binomická regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná s negatívne binomickým rozdelením,

5) Negatívne binomická regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná s negatívne binomickým rozdelením,
- $X = (X_1, \ldots, X_k)$ sú nezávislé vysvetľujúce premenné ovplyvňujúce strednú hodnotu Y,

5) Negatívne binomická regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná s negatívne binomickým rozdelením,
- $X = (X_1, \ldots, X_k)$ sú nezávislé vysvetľujúce premenné ovplyvňujúce strednú hodnotu Y,
- $t(X)$ je offsetová konštanta (čas vystavenia),

5) Negatívne binomická regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná s negatívne binomickým rozdelením,
- $X = (X_1, \ldots, X_k)$ sú nezávislé vysvetľujúce premenné ovplyvňujúce strednú hodnotu Y,
- $t(X)$ je offsetová konštanta (čas vystavenia),
- $\beta = (\beta_1, \ldots, \beta_k)$ sú neznáme regresné koeficienty.

5) Negatívne binomická regresia

Všeobecný model

Nech

- Y je početnostná závislá premenná s negatívne binomickým rozdelením,
- $X = (X_1, \ldots, X_k)$ sú nezávislé vysvetľujúce premenné ovplyvňujúce strednú hodnotu Y,
- $t(X)$ je offsetová konštanta (čas vystavenia),
- $\beta = (\beta_1, \ldots, \beta_k)$ sú neznáme regresné koeficienty.

Všeobecný negatívne binomický regresný model má potom tvar

$$\log E(Y|X) - \log t(X) = \beta' X\,.$$

5) Negatívne binomická regresia

- Ak označíme $\mu_i = t(X_i)\exp\{\beta' X_i\}$, potom log-vierohodnosť v modeli NB2 je

$$\ell(\alpha,\beta) = \sum_{i=1}^{n}\left(\sum_{j=0}^{Y_i-1}\log(\alpha+j)\right) - \log(Y_i!) + Y_i\log\left(\frac{\mu_i}{\alpha}\right) - (Y_i+\alpha)\log\left(1+\frac{\mu_i}{\alpha}\right).$$

5) Negatívne binomická regresia

- Ak označíme $\mu_i = t(X_i)\exp\{\beta' X_i\}$, potom log-vierohodnosť v modeli NB2 je

$$\ell(\alpha,\beta) = \sum_{i=1}^{n}\left(\sum_{j=0}^{Y_i-1}\log(\alpha+j)\right) - \log(Y_i!) + Y_i\log\left(\frac{\mu_i}{\alpha}\right) - (Y_i+\alpha)\log\left(1+\frac{\mu_i}{\alpha}\right).$$

- ML-rovnice sú

$$\sum_{i=1}^{n}\frac{\alpha(Y_i-\mu_i)}{\alpha+\mu_i}X_i = 0,$$

$$\sum_{i=1}^{n}\left[\frac{Y_i-\mu_i}{\alpha+\mu_i} + \log\left(1+\frac{\mu_i}{\alpha}\right) - \sum_{j=0}^{Y_i-1}\frac{1}{\alpha+j}\right] = 0.$$

5) Negatívne binomická regresia

- Asymptotická variančná matica je

$$\operatorname{var}\hat{\beta} = \left(\sum_{i=1}^{n} \frac{\alpha\mu_i}{\alpha+\mu_i} X_i X_i' \right)^{-1},$$

$$\operatorname{var}\hat{\alpha} = \left[\sum_{i=1}^{n} \left(\frac{\mu_i}{\alpha(\alpha+\mu_i)} - \sum_{j=0}^{Y_i-1} \frac{1}{(\alpha+j)^2} \right) \right]^{-1}.$$

5) Negatívne binomická regresia

- Asymptotická variančná matica je

$$\operatorname{var}\hat{\beta} = \left(\sum_{i=1}^{n} \frac{\alpha\mu_i}{\alpha+\mu_i} X_i X_i' \right)^{-1},$$

$$\operatorname{var}\hat{\alpha} = \left[\sum_{i=1}^{n} \left(\frac{\mu_i}{\alpha(\alpha+\mu_i)} - \sum_{j=0}^{Y_i-1} \frac{1}{(\alpha+j)^2} \right) \right]^{-1}.$$

- Použiteľné sú všetky štandardné miery dobrej zhody.

5) Negatívne binomická regresia

- Asymptotická variančná matica je

$$\operatorname{var}\hat{\beta} = \left(\sum_{i=1}^{n} \frac{\alpha\mu_i}{\alpha + \mu_i} X_i X_i' \right)^{-1},$$

$$\operatorname{var}\hat{\alpha} = \left[\sum_{i=1}^{n} \left(\frac{\mu_i}{\alpha(\alpha + \mu_i)} - \sum_{j=0}^{Y_i - 1} \frac{1}{(\alpha + j)^2} \right) \right]^{-1}.$$

- Použiteľné sú všetky štandardné miery dobrej zhody.
- Môžeme testovať existenciu prebytku rozptylu, t.j. či stačí model Poissonovej regresie.

5) Negatívne binomická regresia

- Asymptotická variančná matica je

$$\operatorname{var}\hat{\beta} = \left(\sum_{i=1}^{n} \frac{\alpha\mu_i}{\alpha+\mu_i} X_i X_i' \right)^{-1},$$

$$\operatorname{var}\hat{\alpha} = \left[\sum_{i=1}^{n} \left(\frac{\mu_i}{\alpha\left(\alpha+\mu_i\right)} - \sum_{j=0}^{Y_i-1} \frac{1}{(\alpha+j)^2} \right) \right]^{-1}.$$

- Použiteľné sú všetky štandardné miery dobrej zhody.
- Môžeme testovať existenciu prebytku rozptylu, t.j. či stačí model Poissonovej regresie.
- Parametrizácia modelu používajúca $\alpha^* = 1/\alpha$ je tiež častá.

5) Negatívne binomická regresia

- Asymptotická variančná matica je

$$\operatorname{var}\hat{\beta} = \left(\sum_{i=1}^{n} \frac{\alpha\mu_i}{\alpha+\mu_i} X_i X_i' \right)^{-1},$$

$$\operatorname{var}\hat{\alpha} = \left[\sum_{i=1}^{n} \left(\frac{\mu_i}{\alpha\left(\alpha+\mu_i\right)} - \sum_{j=0}^{Y_i-1} \frac{1}{(\alpha+j)^2} \right) \right]^{-1}.$$

- Použiteľné sú všetky štandardné miery dobrej zhody.
- Môžeme testovať existenciu prebytku rozptylu, t.j. či stačí model Poissonovej regresie.
- Parametrizácia modelu používajúca $\alpha^* = 1/\alpha$ je tiež častá.
- Výsledky pre model NB1 sú podobné.

4) Poissonova regresia

Software dáva pre model NB2 Dollových rakovinových dát ($R^2_{MF} = 0.2767$):

	coefficient	*std. error*	*z*	*P* $>$ |*z*|	*95% conf. interval*	
cigarettes per day	*0.0743164*	*0.0098277*	*7.56*	*0.000*	*0.0550544*	*0.0935784*
years of smoking	*0.1162096*	*0.0107826*	*10.78*	*0.000*	*0.0950761*	*0.1373431*
constant	*-12.60549*	*0.5324672*	*-23.67*	*0.000*	*-13.64911*	*-11.56188*
alpha	*0.2812763*	*0.1323827*			*0.1118188*	*0.7075404*

LR-test hypotézy $H_0: \ \alpha = 0$ dáva $\chi^2(1) = 14.66$, p-hodnota $= 0.000$.

4) Poissonova regresia

Software dáva pre model NB2 Dollových rakovinových dát ($R^2_{MF} = 0.2767$):

	coefficient	*std. error*	*z*	*P* $>$ |*z*|	*95% conf. interval*	
cigarettes per day	*0.0743164*	*0.0098277*	*7.56*	*0.000*	*0.0550544*	*0.0935784*
years of smoking	*0.1162096*	*0.0107826*	*10.78*	*0.000*	*0.0950761*	*0.1373431*
constant	*-12.60549*	*0.5324672*	*-23.67*	*0.000*	*-13.64911*	*-11.56188*
alpha	*0.2812763*	*0.1323827*			*0.1118188*	*0.7075404*

LR-test hypotézy $H_0: \ \alpha = 0$ dáva $\chi^2(1) = 14.66$, p-hodnota $= 0.000$.

Pre model NB1 ($R^2_{MF} = 0.3400$):

	coefficient	*std. error*	*z*	*P* $>$ |*z*|	*95% conf. interval*	
cigarettes per day	*0.0801921*	*0.0073283*	*10.94*	*0.000*	*0.0658288*	*0.0945554*
years of smoking	*0.1050441*	*0.0096732*	*10.86*	*0.000*	*0.0860851*	*0.1240032*
constant	*-12.29042*	*0.4857593*	*-25.30*	*0.000*	*-13.24249*	*-11.33835*
alpha	*0.8527327*	*0.4432867*			*0.3078379*	*2.362129*

LR-test hypotézy $H_0: \ \alpha = 0$ dáva $\chi^2(1) = 8.46$, p-hodnota $= 0.002$.

Utrpenie skončilo...

Ďakujem za pozornosť!

Literature

- D. Collett: Modeling binary data, 2nd ed., Chapman & Hall CRC, 2003
- A. Agresti: Categorical data analysis, 2nd ed., Wiley, 2002
- D.W. Hosmer, S. Lemeshow: Applied logistic regression, 2nd ed., Wiley, 2000
- M. Norusis: SPSS Advanced Statistical Procedures Companion, Prentice Hall, 2005
- J.L. Schafer: Lecture notes for categorial data analysis course, Pensylvania State University, 2006
- D.G. Garson: Statnotes - Topics in Multivariate Analysis, North Carolina State University, 2009
- J.L. Fleiss, B. Levin, M.C. Paik: Statistical Methods for Rates and Proportions, 3rd ed., Wiley, 2003
- SPSS documentation
- A.C. Cameron, P.K. Trivedi: Regression Analysis of Count Data, Cambridge University Press, 1998
- P. McCullagh, J.A. Nelder: Generalized Linear Models, 2nd ed., Chapman & Hall, 1989