

# **ANALÝZA KATEGORIZOVANÝCH DAT V *SOCIOLOGII***

Tomáš Katrňák

Fakulta sociálních studií  
Masarykova univerzita  
Brno

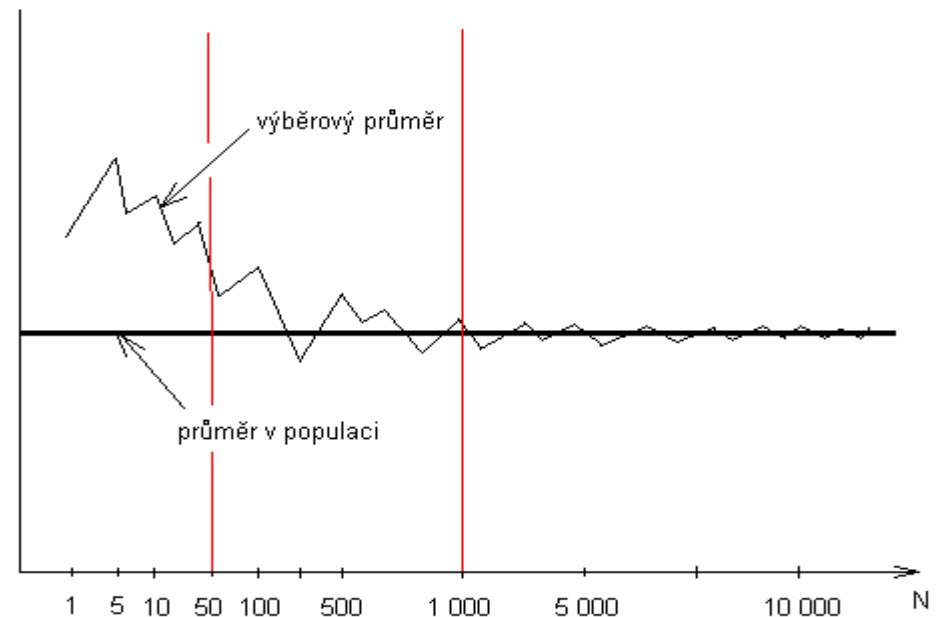
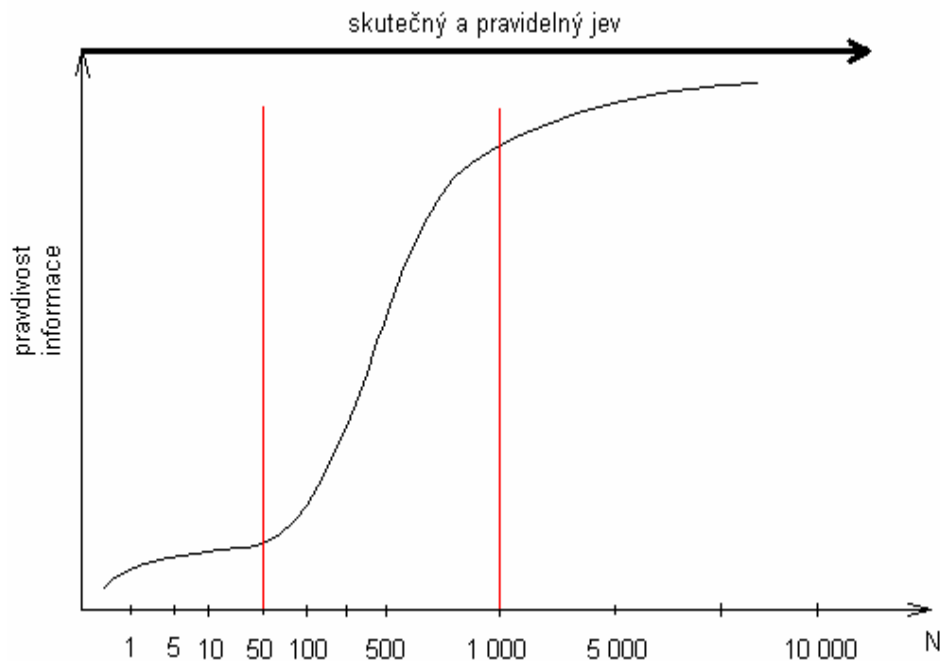
# Lekce 1: Úvod do analýzy kategorizovaných dat

## SOCIOLOGIE A STATISTIKA

- nadindividuální společenské struktury podmiňují lidské chování (Durkheim)
- společenské struktury lze pozorovat na základě statistik o lidském chování, pozorujeme novou skutečnost, z individuálního hlediska nerozpoznatelnou, ptací perspektiva, vymezuje a zároveň přináší informaci o tzv. **hromadném jevu**
  - **hromadný jev** je kolektivita nového řádu, její objevení souvisí s konstitucí moderní společnosti a ustavením sociologie a statistiky jako věd o sociálním životě v moderní společnosti
- **hromadný jev** je definován dostatečným počtem zkoumaných jednotek, protože až na základě určitého počtu (mnohosti) lze získat představu o pravidelnosti, struktuře a zákonitostech v sociálním životě (opakem je **individuální jev**)
  - kde vznikají sociální fakta, když nepřamení z psychiky člověka, ačkoliv jsou její nedílnou součástí? ptá se Durkheim
  - zdroje sociálních faktů leží v sociálních vazbách mezi lidmi, leží tedy v nadindividuálních sociálních strukturách, odpovídá Durkheim
- z tohoto důvodu sociologové pro pochopení sociálního života zkoumají nadindividuální sociální struktury, statistika a statistický aparát jim v tom pomáhají

## ZÁKON VELKÝCH ČÍSEL

- sociální jev je vždy hromadný jev, adjektivum sociální odkazuje k hromadnosti a sociálním vazbám (Simmel)
- všechny jevy (včetně sociálních) podléhají **zákonu velkých čísel** (jako první jej definoval francouzský matematik a statistik Poisson)
  - podle tohoto zákona se empirické údaje o jevu blíží skutečnosti s rostoucím počtem pozorovaných jednotek (když pozorujeme všechny jednotky, pozorujeme skutečnost), pravidelnost a pravá podstata jevu tedy vyvstává na povrch s rostoucím počtem pozorovaných případů



## PROMĚNNÉ A JEJICH DĚLENÍ

- podle slovního vyjádření hodnot proměnných:
  - **kvantitativní proměnné** (diskrétní & spojité)
  - **kvalitativní proměnné**
- podle vztahů mezi hodnotami jednotlivých proměnných:
  - **nominální** (název variant)
  - **ordinální** (název variant + uspořádání vertikální nebo horizontální)
  - **kardinální** (název variant + uspořádání + vzdálenost)
    - **intervalové** (o kolik je jedna hodnota větší než druhá),  $\langle -\infty; \infty \rangle$ , neexistuje racionální 0 (např. teplota ve °C, 0 neznamena nepřítomnost teploty)
    - **poměrové** (kolikrát je jedna hodnota větší než druhá)  $\langle 0; \infty \rangle$ , 0 má racionální základ (např. věk, počet dětí, váha, životnost výrobku atd.)
- hranice mezi jednotlivými proměnnými nejsou neprůchodné, záleží na úhlu pohledu, např. členství v politické straně (nominální, ordinální) nebo vzdělání (nominální, ordinální, kardinální)
- proměnné vyššího řádu měření lze převést do nižšího řádu měření (tzv. ordinalizace nebo nominalizace proměnných)

## PROMĚNNÉ A JEJICH DĚLENÍ

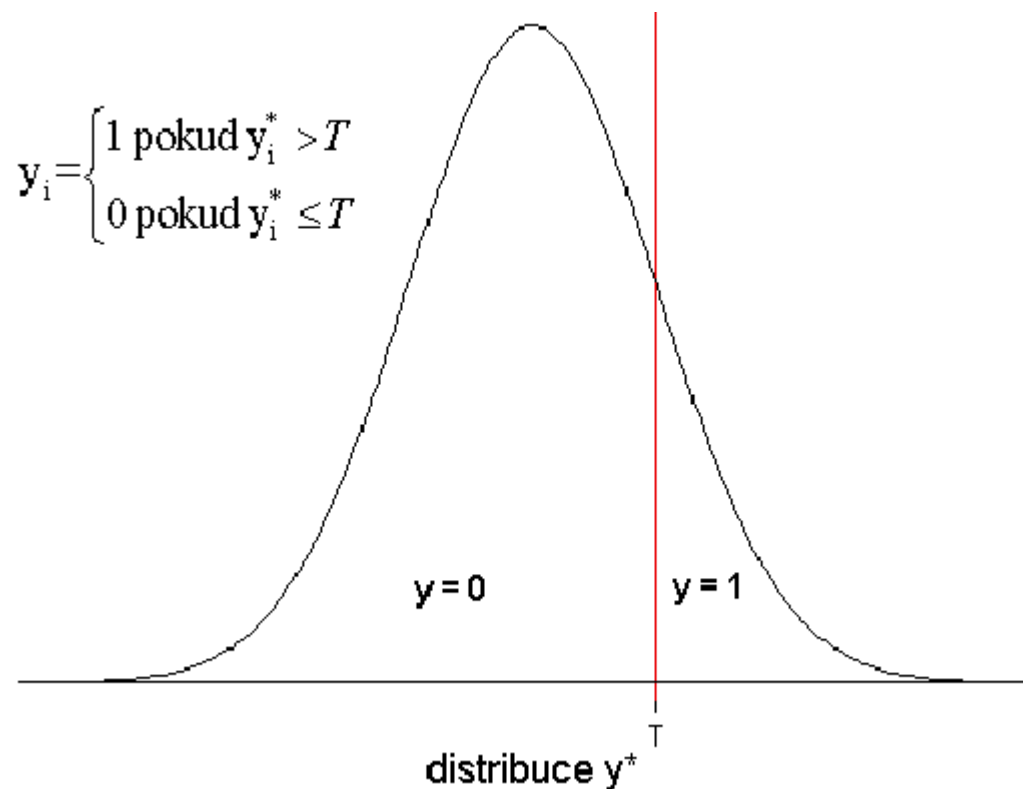
- pod hlavičku kategorizované proměnné řadíme nominální, ordinální a kardinální poměrové proměnné
- kategorizované proměnné dělíme podle počtu variant:
  - **dichotomické** (binární, alternativní)
  - **polytomické** (vícekategoriální)
    - uspořádané kategorie (vertikálně, horizontálně)
    - neuspořádané kategorie (nominální proměnné)

# TRANSFORMAČNÍ PŘÍSTUP VS. PŘÍSTUP LATENTNÍ PROMĚNNÉ

- dva přístupy v konceptualizaci kategorizovaných dat:
  - kategorizovaná data jsou inherentně **diskrétní proměnné** (nelineární pravděpodobnostní modely), statistika a biostatistika, **transformační přístup ke kategorizovaným datům**
    - výběrové varianty znaků a populační parametry se shodují, data modelujeme v měřené podobě,  $p$  (pravděpodobnost) závisle proměnné transformujeme do podoby lineární funkce nezávisle proměnných, tato funkce se nazývá **spojnice** (link), modely transformované do lineárních modelů spojnici se nazývají **zobecněné lineární modely (GLM)** ( $p$  je pak vyjádřena jako nelineární pravděpodobnostní model)
  - kategorizovaná data jsou nepozorované latentní proměnné (hovoří se o modelech latentních proměnných), tento přístup najdeme v ekonometrii a psychometrii, jedná se o tzv. přístup ke kategorizovaným datům jako k latentním proměnným
    - v populaci jsou kategorizovaná data kontinuální proměnné, pozorujeme je ovšem jako kategorizované (např. u proměnné dítě pozorujeme pouze dvě varianty, nicméně v populaci je to kontinuální proměnná, která když překročí určitou míru, tak pozorujeme její manifestaci), rozdíl mezi kontinuálními daty a kategorizovanými daty je v míře pozorovatelnosti, u kategorizovaných dat nevidíme jednotlivé hodnoty, ale pouze intervaly (proto se kategorizovaná data někdy nazývají jako omezené závisle proměnné), v populaci existují tendence, sklon a míra tolerance, přímo úměrné investicím a výnosům, pozorujeme pak jejich projevy skrze překročení míry tolerance

## TRANSFORMAČNÍ PŘÍSTUP VS. PŘÍSTUP LATENTNÍ PROMĚNNÉ

- debata o povaze kategorizovaných dat se táhne historií statistického uvažování (její počátek leží ve sporu K. Pearsona (latentní struktura) a G. U. Yulea (inherentní diskretnost) v první polovině 20. stol., dodnes tato debata není uzavřená, z obou dvou přístupů vycházejí odlišné numerické algoritmy k identifikaci modelů se závisle kategorizovanou proměnnou, jejich výsledky jsou nicméně totožné
- o kterých z následujících proměnných lze uvažovat jako o latentních?
  - úspěch u zkoušky, souhlas s předmanželským sexuálním životem, pohlaví, participace na trhu práce, rodinný stav, přijetí na VŠ, sociální status, gravidita, četba časopisu Respekt, zaměstnanecká mobilita
  - u latentní proměnné  $y^*$  předpokládáme, že je lineárně závislá na pozorované proměnné  $x$ , strukturálním vztahem vyjádřeno:
 
$$y^* = \mathbf{x}_i\beta + \varepsilon_i$$
 nebo pro jednoduchou proměnnou vyjádřeno vztahem
 
$$y^* = \alpha + \beta x_i + \varepsilon_i$$





## INDIVIDUÁLNÍ A AGREGOVANÁ DATA

- **individuální data**

- ukazují varianty znaků pro jednotlivá pozorování
- jednotlivé případy charakterizuje vždy jedna varianta zkoumané proměnné
- data jsou prezentována obvykle ve formě matice, v níž vždy jeden řádek odpovídá jednomu pozorování (případu) a jeden sloupec vždy jedné proměnné (znaku), pole matice pak zachycují varianty proměnných u jednotlivých pozorování (případů)

- **agregovaná data**

- ukazují počet opakujících se pozorování
- jednotlivé kombinace variant proměnných jsou charakterizovány počtem případů
- data jsou prezentována obvykle ve formě kontingenční tabulky, v řádcích a sloupcích tabulky jsou zkombinovány varianty proměnných, v polích tabulky jsou četnosti pozorování (počty případů) těchto variant

## AGREGOVANÁ DATA A JEJICH ANALÝZA POMOCÍ STATY

- **fully relational format of data** - každé pole tabulky odpovídá jednomu pozorování, pole tabulky jsou v matici soustředěné pod jednu proměnnou
- **folded (grouped) format of data** - pozorování je o polovinu méně než polí v tabulce, nicméně pozorování jsou soustředěná pod dvě proměnné (tedy do šířky matice)
- **příklad:**

	Age through 54		Age through 55 and above	
	tolbutamine	placebo	tolbutamine	placebo
Dead	8	5	22	16
Surviving	98	115	76	79

(1) fully relational format

	agecat	exposed	died	pop
1.	0	1	1	8
2.	0	1	0	98
3.	0	0	1	5
4.	0	0	0	115
5.	1	1	1	22
6.	1	1	0	76
7.	1	0	1	16
8.	1	0	0	69

(2) folded format

	agecat	exposed	deaths	pop
1.	0	1	8	106
2.	0	0	5	120
3.	1	1	22	98
4.	1	0	16	85

- podle typu dat volíme ve Statě syntax výpočtu, např. **logit** akceptuje (1), **blogit** akceptuje (2), **glogit** akceptuje (2), ale odhad není proveden jako ML, ale jako WLS, **glm** akceptuje jak (1), tak (2), obecně je ve Statě rozšířenější typ dat (1)

## Lekce 4:

Lineární regresní model, zobecněné lineární modely (GLM), principy statistického modelování

## JEDNODUCHÁ LINEÁRNÍ REGRESE

- jednoduchá lineární regrese může být pro výběrový soubor zapsána jako:

$$\hat{y}_i = a + bx_i + d_i$$

- pro populaci je pak zapsána jako:

$$y_i = \alpha + \beta x_i + e_i$$

- kde  $\alpha$  je posunutí (intercept),  $\beta$  je směrnice pro jednotlivé varianty  $x$  a  $e$  je chyba (residuál, odchylka) pozorované proměnné od odhadnuté směrnice
- v regresní analýze je hodnota závisle proměnné specifikována jako součet lineárních efektů nezávisle proměnné (prediktora) a chyb (residuálů, odchylek, diferencí)

# JEDNODUCHÁ LINEÁRNÍ REGRESE

- stata syntaxt pro regresní model

```
regress price mpg headroom trunk weight length
```

Source	SS	df	MS			
Model	242096575	5	48419315.1	Number of obs =	74	
Residual	392968821	68	5778953.25	F( 5, 68) =	8.38	
Total	635065396	73	8699525.97	Prob > F =	0.0000	
				R-squared =	0.3812	
				Adj R-squared =	0.3357	
				Root MSE =	2403.9	

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mpg	-85.75773	83.60769	-1.03	0.309	-252.5943	81.07882
headroom	-710.1846	444.8546	-1.60	0.115	-1597.878	177.5089
trunk	111.1498	109.9446	1.01	0.316	-108.2411	330.5408
weight	4.420511	1.165629	3.79	0.000	2.094535	6.746488
length	-108.0777	42.56471	-2.54	0.013	-193.0142	-23.1411
_cons	15552.1	6027.182	2.58	0.012	3525.049	27579.16

## ODHAD JEDNODUCHÉ LINEÁRNÍ REGRESE - OLS

- když známe vzorec pro regresi:

$$\hat{Y} = a + bX$$

- tak na základě metody nejmenších čtverců směrnici a posunutí vypočítáme podle vzorců:

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

- směrnice prochází vždy průměry obou proměnných, a proto posunutí (a) vypočítáme dosazením průměrů do upravené regresní rovnice

## PŘEDPOKLADY LINEÁRNĚ REGRESNÍHO MODELU

- odpovídající funkční podoba (linearita a normální rozložení)
- minimální výskyt odlehlých pozorování
- normální rozložení náhodných chyb (residuálů), **problém podoby podmíněné distribuce** (např. podmíněné zešikmení)
- homoskedasticita (konstantní variabilita) náhodných chyb (residuálů, složek), **problém tvaru podmíněné distribuce** (např. podmíněná špičatost)
- neexistence korelace mezi náhodnými chybami (residuály) a vysvětlujícími proměnnými
- neexistence multikolinearity

## STATISTICKÁ INFERENCE V LINEÁRNÍ REGRESI

- konfidenční interval  $CI(\beta)$ .....  $b \pm t^*SE_b$
- test hypotézy  $H_0: \beta = 0$ , výpočet  $t$  statistiky:

$$t = \frac{b}{SE_b}$$

- tabulkové kritické hodnoty  $t$  rozdělení



## ODHAD JEDNODUCHÉ LINEÁRNÍ REGRESE - MLE

- cílem MLE (maximálně věrohodného odhadu) je najít takovou hodnotu koeficientu (parametru), který nejméně pravděpodobně generuje výběrová data
- výběrové hodnoty  $y_i$  jsou výsledkem pravděpodobnostní (hustotní) funkce  $f(y_i|\theta)$ , kde  $\theta$  je neznámý parametr, který generuje hodnoty  $y$  v populaci
- věrohodnostní funkce je pak součin pravděpodobností (hustot) jednotlivých  $y_i$ :

$$L = \prod_{i=1}^n f(y_i; \theta)$$

$$\ln L = \sum_{i=1}^n \ln f(y_i; \theta)$$

- hledáme takový koeficient (obvykle sadu koeficientů) které maximalizují  $L$ , MLE tedy porovnává všechny možné regresní koeficienty a odpovídá na otázku, s jakou věrohodností generují naměřená data, numericky je snazší počítat s přirozeným logaritmem  $L$  (hledáme maximum  $\ln L$ , což odpovídá maximu  $L$ )
- k maximalizaci věrohodnostní funkce je nutné znát matematický vzorec pro náhodný proces generující data v populaci
- v případě lineární regrese musíme tedy přijmout předpoklad o rozložení  $y$  ve variantách  $x$ , (neboli předpokládat distribuci residuálů na základě určitého algoritmu)
- pro spojité znaky v regresi platí, že residuály jsou nezávislé, mají konstantní variabilitu  $\sigma^2$  a normální rozložení s  $\mu=0$ .

## ODHAD JEDNODUCHÉ LINEÁRNÍ REGRESE - ML

- u spojitě závislé proměnné předpokládáme, že je generována na základě normálního rozložení (Gaussova distribuce)

- pravděpodobnost (hustota) je:  $p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$  **pro**  $-\infty < y < \infty$

- střední hodnota je:  $\mu = \alpha + \beta x_i$

- dosazením a pro parametry  $\alpha$  a  $\beta$  dostaneme:  $p(y_i | \alpha, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2\right]$

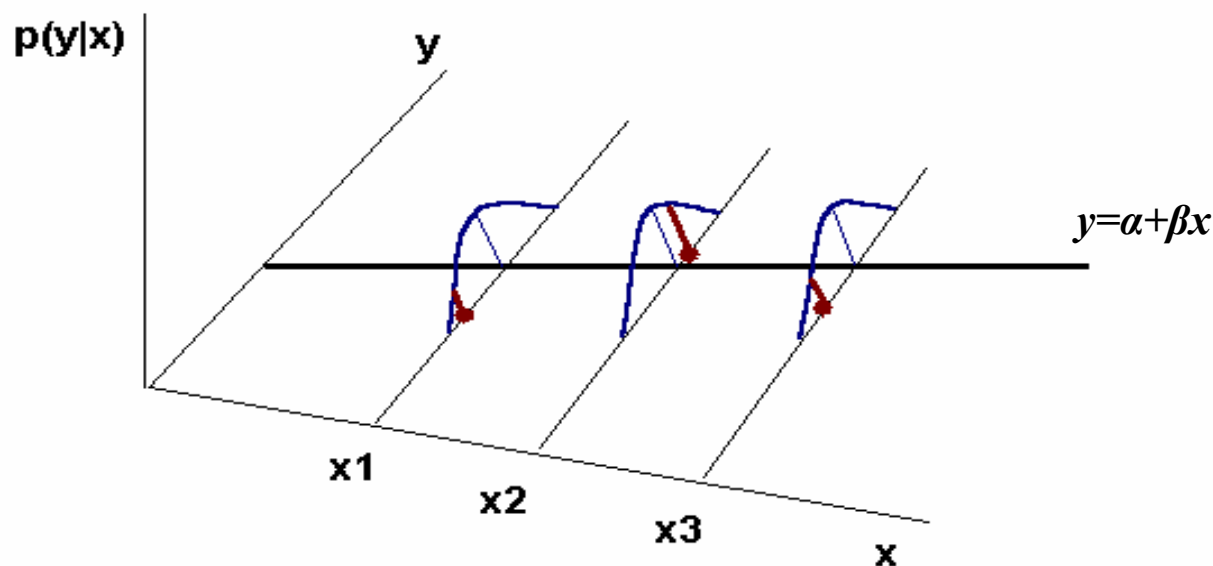
- $y_i$  a  $x_i$  jsou dány, zkoumáme pravděpodobnost pro varianty parametrů  $\alpha$  a  $\beta$
- věrohodnostní funkce  $L$  a přirozený logaritmus věrohodnostní funkce  $\ln L$ :

$$L(\alpha, \beta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2\right]$$

$$\ln L(\alpha, \beta) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

## MLE A ZOBECNĚNÉ LINEÁRNÍ MODELY (GLM)

- $\ln L$  je součet všech pravděpodobností (hustot), která mají jednotlivé hodnoty  $x$  při daných parametrech
- cílem MLE je najít takové parametry, které maximalizují tento součet (je co nejbližší 0), v ideálním případě se tento součet = 0, všechna  $x$  leží na přímce a při daných parametrech mají  $\ln$  hodnoty pravděpodobnosti (hustoty) 0 ( $\exp^0 = 1$ )
- odhad koeficientů na základě MLE je totožný s odhade koeficientů na základě OLS, největší hodnoty  $\ln L$  dostaneme, když je v části vzorce  $(y_i - (\alpha + \beta x_i))^2$  rozdíl minimální ( $y_i$  se co nejvíce blíží  $\mu$ ), volíme tedy takové hodnoty parametrů  $\alpha$  a  $\beta$ , aby to platilo, OLS minimalizuje ten samý vztah, nicméně v termínech residuálů



## MLE A ZOBECNĚNÉ LINEÁRNÍ MODELY (GLM)

- když  $f$  je počet událostí z  $N$  pokusů (tedy pro pravděpodobnost  $y=1$ ) přijímáme předpoklad **binomického rozdělení**; po úpravě pro  $y=0$  (událost nenastala) a  $y=1$  (událost nastala) přijímáme **Bernoulliho rozdělení**
- když  $f$  je počet událostí v čase ( $y=1$ ), v místě nebo v rámci sociální skupiny (neznáme ovšem  $N$ , či počet událostí, které nenastaly ( $y=0$ )), přijímáme předpoklad **Poissonova rozdělení**
- všechna tato rozdělení patří do jedné rodiny distribucí (family), které matematicky vyjadřují náhodný proces, který generuje data (podle jejich typu)
- na základě těchto rozdělení a s pomocí spojnice (link) mezi závisle a nezávisle proměnnou lze tyto případy zobecnit
- hovoříme o zobecněných lineárních modelech (GLM)

## ZOBECNĚNÉ LINEÁRNÍ MODELY

- lineární prediktor  $v_i$  pro každou jednotku je:  $v_i = \mathbf{x}'_i \boldsymbol{\beta}$
- spojnice (*link function*)

$$\text{Identity: } \mu_i = v_i \qquad \text{Logit: } \mu_i = \frac{\exp(v_i)}{1 + \exp(v_i)} \Leftrightarrow \ln\left(\frac{\mu_i}{1 - \mu_i}\right) = v_i$$

$$\text{Log: } \mu_i = \exp(v_i) \Leftrightarrow \ln(\mu_i) = v_i \qquad \text{Probit: } \mu_i = \Phi(v_i) \Leftrightarrow \Phi^{-1}(\mu_i) = v_i$$

- podmíněné distribuce (*exponential family*):

$$\text{Gaussian: } y = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$

$$\text{Binomial: } P(y) = \frac{N!}{y!(N-y)!} \pi^y (1 - \pi)^{N-y}$$

$$\text{Poisson: } P(y) = \frac{e^{-\mu} \mu^y}{y!}$$

**Gamma:**

## ZOBECNĚNÝ LINEÁRNÍ MODEL (GLM)

- jednoduchý regresní model je definován jako strukturní model:

$$Y_i = X_i' \beta + e_i$$

kde  $X_i$  je vektor hodnot pro  $i$ -té pozorování,  $\beta$  je vektor parametrů a  $e$  je chyba.

- statistický model je ve většině případů obsahuje:
  - **fixní část** (*fixed part, systematic component*), která popisuje vztah mezi proměnnými, které nás zajímají (tento vztah je obvykle lineární, a proto umožňuje zodpovědět otázku, jak proměnná  $X$  ovlivňuje  $Y$ )
  - **náhodná část** (*random part, random component*), jedná se o (reziduální) variaci vysvětlované proměnné, která je predikována na základě fixní části

## TYPY ZOBECNĚNÝCH LINEÁRNÍCH MODELŮ

Fixní část	Link	Náhodná část	Model
spojitá	identity	normální	regresní model
kategorizovaná	identity	normální	ANOVA
mix	identity	normální	ANCOVA
mix	logit	binomická	logistická regrese
mix	log	poisson	loglineární analýza
mix	zobecněný logit	multinomická	multinomická logistická regrese

## ZOBECNĚNÉ LINEÁRNÍ MODELY (POKR.)

- stata syntaxt pro GLM

```
glm depvar varlist, family( ) link( )
```

kde

<u>Family</u>	<u>Default</u>	<u>Link(spojnice)</u>	<u>Other link</u>
gaussian	identity	xb	
binomial	logit	$\exp(xb)/(1+\exp(xb))$	probit, c-log-log
poisson	log	$\exp(xb)$	
gamma	log	$\exp(xb)$	1/xb



## CO JE DOBRÝ STATISTICKÝ MODEL?

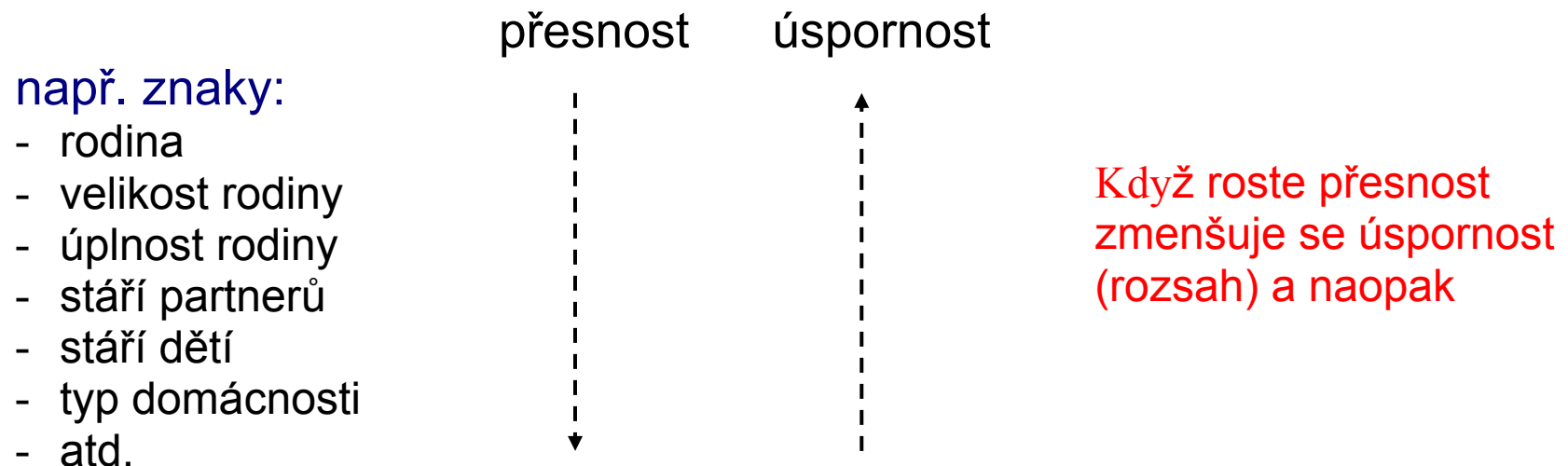
- dobrý statistický model je:
  - je **přesný** (očekávané četnosti, variabilita, podmíněný průměr) se co nejméně liší od pozorovaných četností, variability, podmíněného průměru)
  - je **úsporný** (obsahuje co nejméně parametrů, vysvětlujících proměnných)
    - koncept přesnosti (*accuracy*) = statistická kritéria  $X^2$ ,  $L^2$
    - koncept úspornosti (*parsimony*) = stupně volnosti (d.f. degrees of freedom)
- satureovaný model (obsahuje všechny vysvětlující proměnné a vztahy mezi nimi) **je přesný** (očekávané = pozorované četnosti,  $X^2$  a  $L^2 = 0$ ,  $df = 0$ ), ale **není úsporný** (je to parametrizace pozorovaných četnosti, nic nevysvětluje)
- model (podmíněné) nezávislosti (obsahuje obvykle minimum proměnných a vztahů mezi nimi), **je úsporný**, ale obvykle **není přesný** (rozdíl mezi očekávanými pozorovanými četnostmi je velký,  $X^2$  &  $L^2$  vysoké,  $df$  vysoké, málo parametrů na explanaci)

## PRINCIPY STATISTICKÉHO MODELOVÁNÍ

- v modelování výzkumník obvykle postupuje tak, že hledá model (v případě, že model (podmíněné) nezávislosti na data nepadne), který se nachází někde mezi saturevaným modelem a modelem nezávislosti
- modelování je hledání optimálního poměru mezi **přesností** a **úsporností** (logika Occamovy břitvy)
- cílem je najít co nejúspornější model, který má co nejméně vysvětlujících proměnných, který ovšem stále ještě uspokojivě vysvětluje strukturu dat
- důvod minimalizace vysvětlujících proměnných v modelu
  - numerická stabilita
  - snadná zobecnitelnost a aplikovatelnost
- dva možné postupy statistického modelování
  - začneme saturevaným modelem a postupně vylučujeme proměnné (snižuje se přesnost, ale roste úspornost) (*backward elimination in stepwise regression*)
  - začneme modelem (podmíněné) nezávislosti a postupně přidáváme proměnné (snižuje se úspornost, ale roste přesnost) (*forward addition in stepwise regression*),
  - v obou případech je kritériem pro proměnnou v modelu statistická významnost (obvykle 95%), problém hranice!
- **dobrá teorie je základem pro oprávněnost nebo neoprávněnost proměnných v modelu**

## VZTAH MEZI PŘESNOSTÍ A ÚSPORNOSTÍ V SCLG. VÝZKUMU

- každý zkoumaný (výběrový) soubor je definován **obsahem** a **rozsahem**
  - **obsah**: zkoumaný počet společných **znaků** u jednotek, konkrétnost, přesnost
  - **rozsah**: **počet** jednotek, úspornost
- větší obsah znamená větší počet znaků u jednotky, větší přesnost ve vymezení jednotky, nicméně to znamená vymezení menšího počtu jednotek (maximální počet znaků = 1 jednotka),
- větší rozsah, více zkoumaných jednotek, znamená menší počet znaků u jednotky (maximální rozsah = 1 znak) např. lidé



## REGRESNÍ MODELY PRO KATEG. ZÁVISLE PROMĚNNOU

- v případě kategorizované závisle proměnné regresní model nelze použít
- podle typu závisle proměnné volíme:
  - **binární logistickou regresi** - závisle proměnná má dvě varianty
  - **ordinální logistickou regresi** - závisle proměnná více uspořádaných variant
  - **nominální (multinomickou) logistickou regresi** - závisle proměnná více variant

### Shrnutí jednotlivých typů analýzy:

Závisle proměnná	Nezávisle proměnná	Typ analýzy
spojitá	spojitá	regrese, korelační analýza
spojitá	kategorizovaná	regrese, ANOVA
dichotomická (binární)	kategorizovaná	logit/probit, loglinear
dichotomická (binární)	spojitá	logit/probit
neuspořádaná polytomická	kategorizovaná	loglinear, mlogit
neuspořádaná polytomická	spojitá	mlogit
uspořádaná polytomická	kategorizovaná	ologit/oprobit, loglinear
uspořádaná polytomická	spojitá	ologit/oprobit
tabulková data (poměry)	kategorizovaná	loglinear
censored duration data	spojitá, kategorizovaná	loglinear, logit/log-log

# Lekce 5:

## Modely pro binární závisle proměnnou

## LINEÁRNÍ PRAVDĚPODOBNOSTNÍ MODEL - LPM

- závisle proměnná je kategorizovaná, má dvě varianty (obvykle 0 - jev nenastal, 1 - jev nastal), nezávisle proměnné mohou být jak kategorizované, tak spojité
- klasický regresní model se známými předpoklady je:

$$y_i = \alpha + \beta x_i + e_i$$

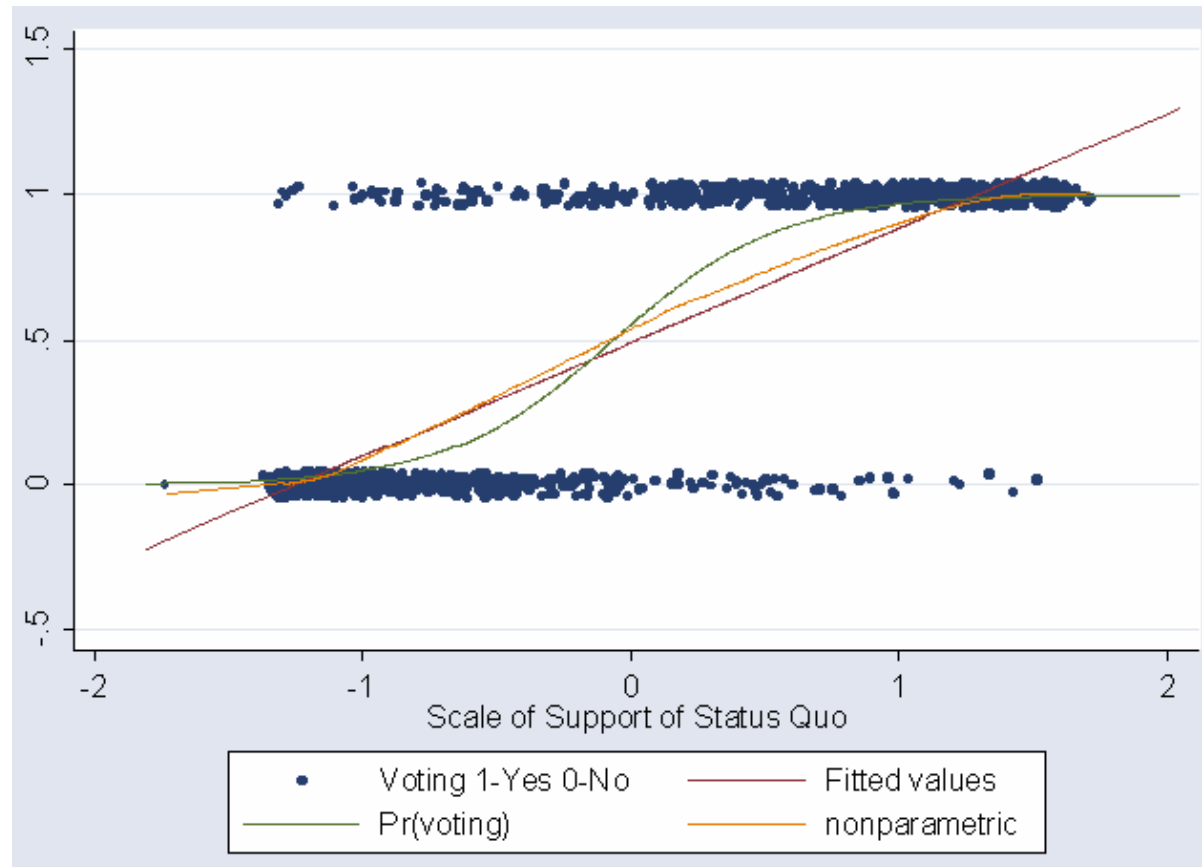
- LPM je modelován jako výskyt podmíněné pravděpodobnosti  $y_i$  při daném  $x_i$
- rovnice modelu je:

$$\Pr(y_i = 1 | x_i) = \pi_i = \alpha + \beta X_i$$

kde očekávaná četnost  $y_i$  při daném  $x_i$  je pravděpodobnost, že  $y_i = 1$  (jev nastal), když je dáno  $x_i$ .

- problémy při identifikaci modelu
  - heteroskedasticita
  - normalita
  - nereálné predikce ( $>1$ ;  $<0$ )
  - funkcionální forma

# FUNKČNÍ ZÁVISLOST VYSVĚTLOVANÉ PROMĚNÉ U LPM



## NELINEÁRNÍ PRAVDĚPODOBNOSTNÍ MODEL (NPM) - LOGIT

- **transformační přístup**
- dvě transformace ve vysvětlované binární proměnné u lineárního pravděpodobnostního modelu před odhadem parametrů
  - první transformace do šancí, podmínka splňuje, že predikované hodnoty budou v intervalu  $\langle 0; \infty \rangle$ ;
 
$$\frac{\Pr(y = 1 | x)}{\Pr(y = 0 | x)} = \frac{\pi}{1 - \pi}$$
  - druhá transformace na přirozený logaritmus šancí, podmínka splňuje, že šance se nacházejí v intervalu  $\langle -\infty; \infty \rangle$ 

$$\ln \left[ \frac{\pi}{1 - \pi} \right]$$
- přirozený logaritmus šancí je nazván v teorii GLM jako **LOGIT** a model je **lineární**, ovšem v transformované (logitové) podobě pro  $\Pr(y=1)$ , a **nelineární** pro pravděpodobnost  $\Pr(y=1)$ , hovoříme pak o **nelineárním pravděpodobnostním modelu (NPM)**



## NPM-LOGIT MODEL

- rovnice logistické regrese (model je lineární jako logit)

$$\ln \left[ \frac{\pi_i}{1 - \pi_i} \right] = \alpha + \beta X_i$$

- pravděpodobnost  $\Pr(y=1)$  je ovšem již na základě takto specifikovaného modelu nelineární

$$\pi_i = \frac{\exp^{(\alpha + \beta X_i)}}{1 + \exp^{(\alpha + \beta X_i)}}$$

- distribuce chyb
- stata syntax odhadu binárního logitového modelu v GLM  
`glm depvar varlist, family(binomial) link(logit)`
- stata syntax odhadu binárního logitového modelu  
`logit depvar varlist`  
`logistic depvar varlist`

## NPM-PROBIT MODEL

- cdf (kumulativní distribuční funkce) splňuje požadavek rozmezí pravděpodobnosti  $\langle 0; 1 \rangle$ , transformací závisle proměnné do této podoby dostaneme probitovou regresi (model je lineární jako probit)

$$\left[ \int_{-\infty}^{\alpha + \beta X_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \right]^{-1} \pi_i = \Phi^{-1} \pi_i = \alpha + \beta X_i$$

- pravděpodobnost  $\Pr(y=1)$  je ovšem již na základě takto specifikovaného modelu nelineární

$$\pi_i = \int_{-\infty}^{\alpha + \beta X_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt = \Phi(\alpha + \beta X_i)$$

- distribuce chyb
- stata syntax odhadu binárního probitového modelu v GLM  
`glm depvar varlist, family(binomial) link(probit)`
- stata syntax odhadu binárního probitového modelu  
`probit depvar varlist`

## NPM - KOMPLEMENTÁRNÍ LOG-LOG MODEL

- komplementárního log-log modelu je další variantou transformace závisle proměnné, které je pak lineárním vyjádřením parametrů:

$$\ln(-\ln[1 - \pi_i]) = \alpha + \beta X_i$$

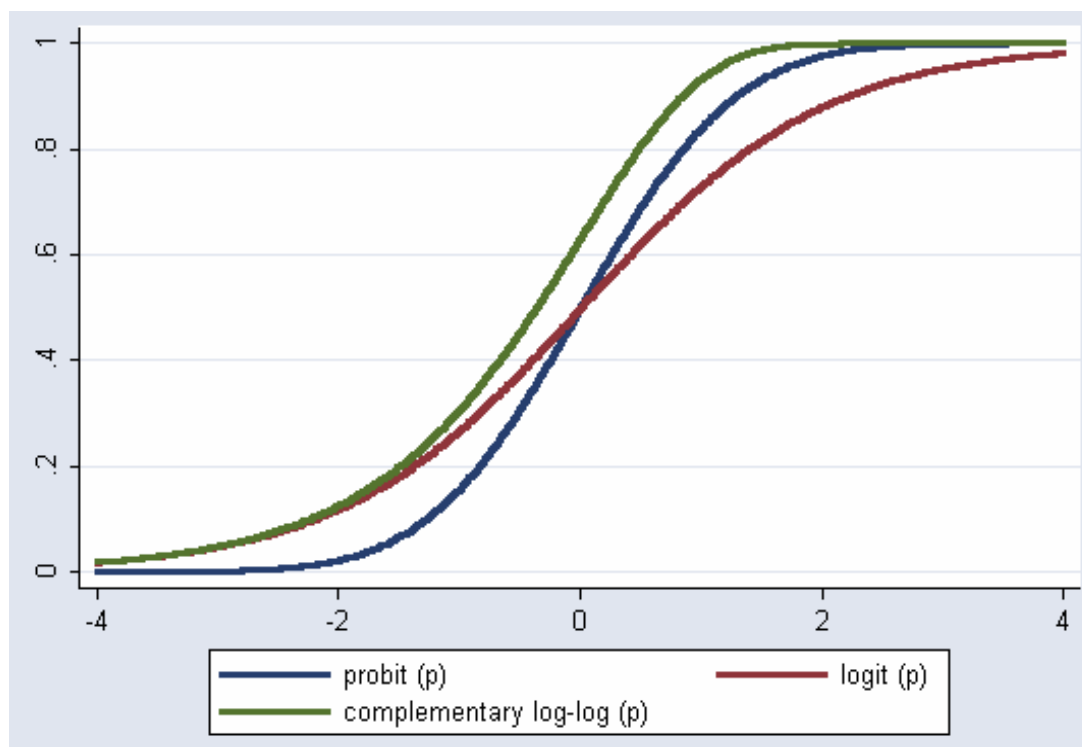
- pravděpodobnost  $y=1$  je pak vyjádřena jako nelineární funkce:

$$\pi_i = 1 - \exp[-\exp(\alpha + \beta X_i)]$$

- distribuce chyb
- stata syntax odhadu binárního komplementárního log-log modelu  
`cloglog depvar varlist`

## DISTRIBUČNÍ FUNKCE LOGIT, PROBIT A LOG-LOG MODELU

- predikované hodnoty  $\Pr(y=1|x)$  podle logitového, probitového a komplementárního log-log modelu, logit a probit podobné, kompl. log-log model dává substantivně odlišné výsledky



## MAXIMÁLNĚ VĚROHODNÝ ODHAD (MLE)

- binomická **pravděpodobnostní funkce** pro  $y$  úspěchů, při pravděpodobnosti na úspěch  $\pi$ , v  $N$  pokusech je:

$$f(\pi) = \Pr(y | n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad \binom{n}{y} = \frac{n!}{y!(n-y)!}$$

$$E(y) = n\pi \quad \text{Var}(y) = n\pi(1 - \pi)$$

- známe matematický vzorec pro určení pravděpodobnosti (vzorec pro náhodný proces, který generuje data) a chceme znát pravděpodobnost určitého výsledku (např. 3 mužů, ve vzorku  $n=10$ , při  $\pi=0.5$ )
- typický problém: ve statistice známe výsledek  $y$  a  $n$ , neznáme ovšem parametr  $\pi$ , který musíme z informací ve výběru odhadnout
- binomická **věrohodnostní funkce** je:

$$L = \prod_i f(\pi_i) = \prod_i \Pr(\pi_i | y, n) = \prod_i \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

- věrohodnostní funkce ukazuje, jak je pravděpodobné, že budeme pozorovat data, která pozorujeme při hodnotách určitých parametrů
- maximálně věrohodný odhad je potom taková hodnota parametru, která s nejvyšší pravděpodobností (nejvěrohodněji) generuje pozorovaná data

## ODHAD LOGITOVÉHO MODELU (WLS, MLE)

- WLS (odhad pomocí Weighted least square), glogit (používá se velmi zřídka)
- MLE je nezbytné použít, protože efekt nezávisle proměnných na závisle proměnnou není lineární, residuály nemají normální distribuci a pro hodnoty nezávisle proměnné není jejich variance konstantní (glm, logit, logistic)
- cílem MLE je nalézt koeficienty nezávisle proměnných, které generují data, jež co nejvíce odpovídají pozorovaným datům, to lze provést pomocí maximalizace věrohodnostní funkce;

$$L = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{(1-y_i)}$$

binomický koeficient

$\binom{n_i}{y_i}$  není nutné v

kde  $L$  je hodnota věrohodnostní funkce;  $p_i$  je predikovaná pravděpodobnost pro případ  $i$  podle vzorce  $p_i = e^{\text{LOGIT}} / (1 + e^{\text{LOGIT}})$ ;  $y_i$  je hodnota nezávisle proměnné pro případ  $i$ ,  $\Pi$  je multiplikativní ekvivalent  $\Sigma$  (funkce je výsledkem násobení hodnot pro každý případ)

rovnici použít, protože pouze konstantně násobí odhad parametrů

- klíčové je identifikovat  $\beta$  koeficienty nezávisle proměnných, které produkují LOGIT a zároveň tak  $p$ , čím maximalizují  $L$

$$L = \prod_{i=1}^n F(x_i' \beta)^{y_i} [1 - F(x_i' \beta)]^{(1-y_i)}$$

## ODHAD LOGITOVÉHO MODELU (MLE)

- numericky je ovšem snazší pracovat s přirozeným logaritmem věrohodnostní funkce (vyhneme se multiplikaci pravděpodobností a extrémně nízkým kladným číslům)
- když věrohodnostní funkce maximalizuje pravděpodobnost, tak její přirozený logaritmus maximalizuje přirozený logaritmus pravděpodobnosti

$$L = \ln L = \sum_{i=1}^n \{y_i \ln [\pi(x_i)] + (1 - y_i) \ln [1 - \pi(x_i)]\}$$

- zatímco  $L$  variuje  $\langle 0;1 \rangle$ ;  $\ln L$  variuje  $\langle -\infty;0 \rangle$ , čím blíže je  $L$  1 nebo čím blíže je  $\ln L$  0, s tím větší věrohodností parametry modelu generují pozorovaná data, jedná se o maximalizaci věrohodnostní funkce nebo o maximalizaci přirozeného logaritmu věrohodnostní funkce

$$L = \ln L = \sum_{i=1}^n \{y_i \ln F(\mathbf{x}'_i \boldsymbol{\beta}) + (1 - y_i) \ln [1 - F(\mathbf{x}'_i \boldsymbol{\beta})]\}$$

## OBECNÝ POSTUP PŘI ODHADU POMOCÍ MLE

- 1) volba koeficientů u nezávisle proměnných, 1 iterace obsahuje obvykle pouze  $\alpha$  (např. ve Statě) nebo koeficienty odhadnuté na základě OLS, další varianty koeficientů se již vybírají na základě iterací
  - 2) výpočet predikovaného LOGITU na základě zvolených koeficientů  $\alpha$ ,  $\beta$  a případu  $x_i$
  - 3) transformace LOGITU do pravděpodobnosti  $p_i$  podle vzorce  $p_i = e^{\text{LOGIT}} / (1 + e^{\text{LOGIT}})$
  - 4) výpočet přirozeného logaritmu hodnoty věrohodnostní funkce pro případ  $x_i$
  - 5) opakujeme krok 1 až 4 pro všechny případy  $x_i$ , sečteme a dostaneme tak hodnotu **přirozeného logaritmu věrohodnostní funkce ( $\ln L$ )** pro zvolené koeficienty
  - 6) opakujeme kroky 1 až 5 pro všechny možné varianty kombinací koeficientů a srovnáváme jejich  $\ln L$
  - 7) volíme tu variantu kombinace koeficientů, která má nejvyšší hodnotu  $\ln L$  (nejblíže 0)
- **konečná hodnota  $\ln L$  ukazuje míru věrohodnosti, že dostaneme pozorovaná data, při daných koeficientech nezávisle proměnných (parametrech)**



## MLE LOGIT MODELU V KONTINGENČNÍ TABULCE

- věrohodnostní funkce

$$L = \prod_i \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

- přirozený logaritmus věrohodnostní funkce

$$\ln L = \sum_{i=1}^n \{y_i \ln [\pi(x_i)] + (n_i - y_i) \ln [1 - \pi(x_i)]\}$$

$$\ln L = \sum_{i=1}^n \{y_i \ln F(\mathbf{x}'_i \boldsymbol{\beta}) + (n_i - y_i) \ln [1 - F(\mathbf{x}'_i \boldsymbol{\beta})]\}$$

## STATISTICKÁ INFERENCE NPM

- podobné jako v regresní analýze (OLS)
- podíl koeficientů nezávisle proměnných a standardní chyby (SE) je základem testu významnosti (z distribuce)

$$z = \frac{\beta}{SE}$$

- statistická významnost koeficientů označuje pravděpodobnost, že velikost výběrových koeficientů je náhodná, když populační parametry odpovídají 0, v sociálních vědách si obvykle necháváme rezervu 5% pro náhodu
- pro spolehlivost testu významnosti by  $N > 100$

## KOMPLEXNĚJŠÍ TESTY VÝZNAMNOSTI – WALDŮV TEST

- oboustranný test významnosti jednotlivých koeficientů nebo jejich simultánního efektu
- Waldův test je umocněná  $t$ -statistika ( $t$ -ratio) a odpovídá chí-kvadrát distribuci rozdělení pravděpodobností

$$W = \left( \frac{\beta_x}{SE} \right)^2$$

- test jednoduché nulové hypotézy ( $\beta_1 = 0$ ),
- test komplexnější nulové hypotézy ( $\beta_1 = \beta_2 = 0$ ) nebo ( $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ )
- Waldův test je postaven na omezování jednoho nebo více odhadnutých parametrů v jednom odhadnutém modelu (testujeme např. nulovou hypotézu, podle níž efekt  $\beta_1 = 0$  a  $\beta_2 = 0$ , nebo nulovou hypotézu, podle níž  $\beta_1 = \beta_2$ ), když  $p \geq 0.05$ , tak toto omezení není náhodné
- u Waldova testu není nutné odhadovat 2 modely, jako je tomu u lrtestu

- **stata syntax**

```
.logit chd age age2 sex
.test age2
.test age2 age
.test age2=age
```

## KOMPLEXNĚJŠÍ TESTY VÝZNAMNOSTI – LRTEST

- test významnosti komplexnějších hypotéz o odhadnutých koeficientech (např. test významnosti simultánního efektu více regresorů)
- základem je porovnání dvou maximalizovaných hodnot věrohodnostních funkcí z různých modelů
- **notace:**  $M_0 \ln(L_0)$ ... **základní model** (*baseline model*), neúspornější, obsahuje pouze konstantu (predikovanou průměrnou pravděpodobnost pro všechny případy),  $L_0$  je nejnižší
  - $M_F \ln(L_F)$ ... **plný (navržený) model** (*full model*), přesnější než  $M_0$ , obsahuje pouze konstantu + další koeficienty,  $L_F$  je vždy vyšší než  $L_0$ , protože se jedná o krok k přesnosti
  - $M_s \ln(L_s)$ ... **saturovaný model**, nejpresnější model, úspornost nejnižší, obsahuje všechny možné koeficienty a varianty vztahů mezi nimi,  $L_s = 0$
- hodnota  $\ln(L)$  ukazuje  $\ln$  věrohodnosti, s níž naměříme data při daných koeficientech (čím blíže 0, tím větší věrohodnost), je to tedy **odchylka** od saturovaného modelu
- hodnota  $\ln(L)$  závisí na  $N$  - čím vyšší  $N$ , tím nižší  $\ln(L)$  - a počtu parametrů, posoudit její velikosti je proto nutné skrze standardizovaný algoritmus
- tím je test poměru maximální věrohodnosti (*likelihood ratio test*), krátce LRTEST (v loglineárním modelování  $L^2$  někdy také  $G^2$ )

## KOMPLEXNĚJŠÍ TESTY VÝZNAMNOSTI – LRTEST

$$LR = 2 \ln \left( \frac{L_F}{L_O} \right) = -2 \left( \frac{L_O}{L_F} \right) \quad \begin{array}{l} L_F \text{ je vždy větší než } L_O, \text{ má více koeficientů, je blíže} \\ \text{saturovanému modelu} \end{array}$$

$$LR = 2(\ln L_F - \ln L_O) = -2(\ln L_O - \ln L_F)$$

- násobíme 2 nebo  $-2$ , dostaneme tak hodnotu chí-kvadrátu s *d.f.* [ $df = df(M_F) - df(M_O)$ ], které odpovídají počtu nezávisle proměnných, srovnání této hodnoty s tabulkovou hodnotou  $X^2$  rozdělení testuje nulovou hypotézu, že všechny koeficienty s výjimkou konstanty se rovnají 0 (změna v hodnotě  $L$  vyvolaná nezávisle proměnnými je náhodná a zlepšení se signifikantně neliší od 0), když  $p \geq 0.05$  podpoříme nulovou hypotézu, dva modely se od sebe signifikantně neliší, úspornější model je vhodnější
- stejnou logiku aplikujeme na porovnání jakýchkoliv dvou modelů a testujeme významnost změn v ( $L$ ) podle jednotlivých nezávisle proměnných, jimiž se modely od sebe odlišují

## KOMPLEXNĚJŠÍ TESTY VÝZNAMNOSTI – LRTEST

- test jednoduché nulové hypotézy ( $\beta_1 = 0$ ),
- test komplexnější nulové hypotézy ( $\beta_1 = \beta_2 = 0$ ) nebo ( $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ )
- základem lrtestu je srovnání ( $L$ ) (komplexního, plného modelu) s ( $L$ ) omezeného modelu, který je v něm „vsazen“ (*nested model*), testujeme významnosti **vynechaných** parametrů (d.f. = počet testovaných parametrů)
- **např.**
  - model 1 je „vsazen“ v modelu 3 (test nulové hypotézy  $\beta_2 = \beta_3 = 0$ ),
  - model 2 je vsazen v modelu 3 (test  $\beta_2 = 0$ ),
  - model 2 je vsazen v modelu 3 (test  $\beta_2 = 0$ )
- celkové  $N$  při lrtestu musí být pro všechny modely stejné, jinak je lrtest neplatný

$$\text{model 1: } \text{logit}(p_i) = \beta_0 + \beta_1 x_1$$

$$\text{model 2: } \text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\text{model 3: } \text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- **stata syntax**

```
.logit chd age age2 sex
.est store A
.logit chd age sex
.lrtest A
.est store B
.logit chd sex
.est store C
.lrtest B
.lrtest C A, stats
```

rozhodnutí pro Waldův test nebo lrtest je otázkou konvence, neexistuje racionální argument pro jeden z nich, při velkých souborech dostaneme stejné výsledky, většina statistiků preferuje lrtest, i když při jeho použití musíme odhadovat 2 modely

## TEST SEDNUTÍ MODELU NA DATA (KOMPARACE MODELŮ)

- míry sednutí modelu na data indikují adekvátnost modelu pro popis struktury dat
- měr je několik, nicméně obecně platí, že neexistuje racionální evidence pro to, že padnutí/nepadnutí modelu na data podle jedné míry je optimálnější než padnutí/nepadnutí modelu na data podle jiné míry
- míry padnutí modelu na data musíme vždy používat v kontextu teorie a hypotéz, které testujeme, zvoleného typu analýzy, předchozího výzkumu na dané téma a závěrů, které přinesl, a především vysvětlujících proměnných, jež používáme jako prediktory
- příkaz *fitstat* počítá velké množství statistik testujících padnutí modelu na data

- **stata syntax**

```
logit lfp k5 k618 age wc hc lwg inc
fitstat
logit lfp k5 k618 age wc hc lwg inc
fitstat, saving(mod1)
logit lfp k5 k618 age age2 wc hc lwg inc
fitstat, using(mod1)
```

## SEDNUTÍ MODELU NA DATA (KOMPARACE MODELŮ)

- většina měr sednutí modelu na data vychází z maximální hodnoty věrohodnostní funkce pro daný model

$$L_O \longleftrightarrow L_F \longleftrightarrow L_S$$

- LRTEST je komparace  $L_F$  a  $L_O$ , jak se naměřený model liší od nulového;  $df$  = počet proměnných, které obsahuje  $L_F$  na rozdíl od  $L_O$ , parametry zde přidávány (adding of parameteres)
- $D$  – odchylka je komparace  $L_F$  a  $L_S$ , jak se naměřený model liší od saturovaného modelu:

$$D = -2(\ln L_F - \ln L_S) = -2 \ln L_F$$

$df$ =počet případů minus počet proměnných, parametry jsou zde ubírány (making of constraints in parameteres)

- platí vztah:

$$LR_O - LR_F = D_F - D_O$$



## MÍRY SEDNUTÍ MODELU NA DATA

	regress	logistic logit probit	cloglog	ologit oprobit	clogit mlogit	cnreg intreg tobit	gologit nbreg poisson zinb zip
Log-likelihood	■	■	■ <sup>1</sup>	■	■	■	■ <sup>2</sup>
Deviance & LR chi-square	■	■	■	■	■	■	■
AIC, AIC*n, BIC, BIC'	■	■	■	■	■	■	■
$R^2$ & Adjusted $R^2$	■	□	□	□	□	□	□
Efron's $R^2$	□	■	■	□	□	□	□
McFadden's, ML, C&U's $R^2$	□	■	■	■	■	■	■
Count & Adjusted Count $R^2$	□	■	■	■	■ <sup>3</sup>	□	□
Var(e), Var( $y^*$ ) and M&Z's $R^2$	□	■	□	■	□	■	□

pramen: Long, Freese (2001)

## MÍRY SEDNUTÍ MODELU NA DATA - VARIANTY $R^2$

- **pseudo  $R^2$** , neboli **McFaddenovo  $R^2$** , či také někdy **index věrohodnostního poměru** ukazuje zlepšení v  $\ln L_F$  vzhledem k  $\ln L_O$ , nabývá hodnot  $<0;1>$ , nevysvětluje ovšem variaci v závisle proměnné, která je dána nezávisle proměnnými, protože  $\ln L$  není o variaci definované jako suma  $\varepsilon^2$

$$\text{pseudo } R^2 = 1 - \frac{\ln L_F}{\ln L_O}$$

- další varianty koeficientu determinace:  $R^2$  maximální věrohosnoti; Craggovo & Uhlerovo  $R^2$ , Efronovo  $R^2$
- frekvenční (*count*) a adjustované frekvenční  $R^2$  ukazuje srovnání pozorovaných dat a na základě modelu predikovaných dat (příkaz `lstat` ve statě), ukazuje chybu s jakou je model predikován

# FREKVENČNÍ A ADJUSTOVANÉ FREKVENČNÍ $R^2$

**lstat**

**Logistic model for lfp**

Classified	True		Total
	D	~D	
+	342	145	487
-	86	180	266
Total	428	325	753

Classified + if predicted Pr(D) >= .5

True D defined as lfp != 0

Sensitivity	Pr( +   D)	79.91%
Specificity	Pr( -   ~D)	55.38%
Positive predictive value	Pr( D   +)	70.23%
Negative predictive value	Pr( ~D   -)	67.67%

False + rate for true ~D	Pr( +   ~D)	44.62%
False - rate for true D	Pr( -   D)	20.09%
False + rate for classified +	Pr( ~D   +)	29.77%
False - rate for classified -	Pr( D   -)	32.33%

**Correctly classified 69.32%**

$$R_{count}^2 = \frac{\sum_j n_{jj}}{N}$$

$$R_{Adj\ count}^2 = \frac{\sum_j n_{jj} - \max_r(n_{r+})}{N - \max_r(n_{r+})}$$

- kde  $n_{jj}$  je počet správných predikcí na základě modelu pro výsledek  $j$
- kde  $n_{r+}$  je řádková četnost pro řádek  $r$

## MÍRY SEDNUTÍ MODELU NA DATA - INFORMAČNÍ KRITÉRIA

- účelem informačních kritérií není určit, který model je pravdivější, ale který model podává bohatší informaci o reálném světě, který model má větší vypovídací schopnost o realitě
  - **AIC** (Akaikeovské informační kritérium) (Akaike, 1987)

$$AIC = (-2 \ln L_F + 2P) / N \quad \text{kde } P = \text{počet parametrů (regresorů)} + 1$$

- **BIC** (Bayesovské informační kritérium) (Schwartz, 1978; Raftery, 1986, 1995)

$$BIC = D - df_D \ln N \quad BIC' = LR^2 - df_{LR} \ln N$$

- čím negativnější velikost BIC (čím větší zápornější číslo), tím více informací model přináší o realitě, obecně platí, že je-li  $BIC > 0$ , souvislost v datech není a platí saturovaný model
- tyto statistiky upřednostňují úspornost před přesností, platí:

$$BIC_1 - BIC_2 = BIC'_1 - BIC'_2$$