7.2 Binární logistická regrese

Závisle proměnná je binární, nabývá jen dvou hodnot 0 a 1. Do tohoto tvaru je potřeba závisle proměnnou upravit, Stata totiž bere nulovou hodnotu jako nulu a jakoukoliv nenulovou hodnotu jako jedničku. Následně se používá stejně jako lineární regrese, jen místo příkazu REGRESS použijeme LOGIT. Samotné koeficienty se interpretují příliš složitě, lepší je interpretovat poměry šancí (odds ratio). Ty dostaneme pomocí parametru OR nebo použití příkazu LOGISTIC namísto příkazu LOGIT. Poměr šancí pak slouží jako multiplikativní konstanta (kolikrát vyšší šanci máme, že nastane vysvětlovaný jev, tedy že závisle proměnná nabývá hodnoty 1).

Výsledné poměry šancí pak interpretujeme buď jako zvýšení/snížení šance, nebo jako procentní nárůst. Např. poměr šancí s hodnotou 1,02 znamená, že pokud se nezávisle proměnná zvýší o jednotku (např. jeden rok), zvýší se pravděpodobnost, že jev nastane přesně 1,02×, neboli o 2 %.

```
recode gearn (1 2 = 1) (3/5 = 0), gen (muz_chlebodarce)
logit muz_chlebodarce i.sex age i.educ
logit muz_chlebodarce i.sex age i.educ, or
logistic muz_chlebodarce i.sex age i.educ
```

```
. logistic muz_chlebodarce i.sex age i.educ

Logistic regression                             Number of obs   =       7098
                                                LR chi2(12)     =     319.13
                                                Prob > chi2     =     0.0000
Log likelihood = -4386.3027                     Pseudo R2       =     0.0351

------------------------------------------------------------------------------
muz_chlebodarce | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
          2.sex |   .7193393   .037664    -6.29   0.000     .6491806    .7970802
            age |   1.001877   .0015825     1.19   0.235     .9987802    1.004983
                |
          educ |
              2 |   .7486893   .6882587    -0.31   0.753     .1235393    4.537308
              3 |   .4500209   .4125754    -0.87   0.384     .0746205    2.713983
              4 |   .4412488   .4056129    -0.89   0.373     .0728159    2.673873
              5 |   .4121165    .383097    -0.95   0.340     .0666433    2.548494
              6 |   .2749535   .2520112    -1.41   0.159     .0456124    1.657433
              7 |   .2411588    .222242    -1.54   0.123      .039616    1.468031
              8 |   .2448166    .228142    -1.51   0.131     .0394108    1.520781
              9 |   .2227436   .2066044    -1.62   0.105     .0361642    1.371929
             10 |   .1731962   .1590227    -1.91   0.056     .0286414    1.047326
             11 |    .15047   .1421647    -2.00   0.045     .0236174    .9586686
                |
          _cons |   1.748869   1.610496     0.61   0.544     .2876761    10.63189
------------------------------------------------------------------------------
```

Obrázek 1 Ukázka binární logistické regrese

7.3 Ordinální logistická regrese

Závisle proměnná je ordinální, nabývá několika kategorizovaných hodnot, které lze seřadit, a jsou od sebe stejně vzdáleny. Následně se používá stejně jako lineární regrese, jen místo příkazu REGRESS použijeme OLOGIT. Opět interpretujeme poměry šancí, které dostaneme pomocí parametru OR.

```
ologit gearn i.sex age i.educ, or
```

```
. ologit gearn i.sex age i.educ, or

Iteration 0:   log likelihood = -11269.737
Iteration 1:   log likelihood = -11054.469
Iteration 2:   log likelihood = -11053.714
Iteration 3:   log likelihood = -11053.714

Ordered logistic regression                     Number of obs   =       7098
                                                LR chi2(12)     =     432.05
                                                Prob > chi2     =     0.0000
Log likelihood = -11053.714                     Pseudo R2       =     0.0192

------------------------------------------------------------------------------
       gearn | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.sex |   1.460922   .0634144     8.73   0.000     1.341772    1.590652
         age |   1.000006   .0013158     0.00   0.996     .9974304    1.002588
             |
       educ |
           2 |   1.658338   1.378037     0.61   0.543     .3253464    8.452793
           3 |   2.820823   2.336415     1.25   0.211     .5563491    14.30224
           4 |   2.969407   2.465969     1.31   0.190     .5831526    15.12019
           5 |   3.462534   2.904156     1.48   0.139     .6690517     17.9196
           6 |   4.561151   3.776682     1.83   0.067     .9000581    23.11418
           7 |   5.342975    4.44276     2.02   0.044     1.047121    27.26273
           8 |    5.03423   4.220006     1.93   0.054     .9736495    26.02936
           9 |   5.902391   4.930382     2.13   0.034     1.148161    30.34263
          10 |   6.957424    5.76646     2.34   0.019     1.370739    35.31362
          11 |   8.102759   6.839523     2.48   0.013     1.549299    42.37703
------------------------------------------------------------------------------
```

Obrázek 2 Ukázka ordinální logistické regrese

7.4 Multinomická logistická regrese

Závisle proměnná je ordinální, nabývá několika kategorizovaných hodnot, které lze seřadit, ale – na rozdíl od ordinální logistické regrese – nejsou od sebe stejně vzdáleny. Následně se používá stejně jako lineární regrese, jen místo příkazu REGRESS použijeme MLOGIT. Opět interpretujeme poměry šancí (v tomto případě se ovšem statisticky korektně označují jako relative risk ratio), které dostaneme pomocí parametru RRR.

```
mlogit gearn i.sex age, brr
```

```
. mlogit gearn i.sex age, rrr

Iteration 0:   log likelihood = -11269.737
Iteration 1:   log likelihood = -11196.666
Iteration 2:   log likelihood = -11196.367
Iteration 3:   log likelihood = -11196.367

Multinomial logistic regression                 Number of obs   =       7098
                                                LR chi2(8)      =     146.74
                                                Prob > chi2     =     0.0000
Log likelihood = -11196.367                     Pseudo R2       =     0.0065

------------------------------------------------------------------------------
       gearn |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Rozhodně_souhlasím |
         2.sex |   .6829551   .0564707    -4.61   0.000     .5807775     .803109
           age |   1.015886   .0025401     6.30   0.000      1.01092    1.020877
         _cons |    .287474   .0419038    -8.55   0.000     .2160343     .382538
-------------+----------------------------------------------------------------
Spíše_souhlasím |
         2.sex |   .7022041   .0502813    -4.94   0.000     .6102574    .8080043
           age |   .9974921   .0021376    -1.17   0.241     .9933111    1.001691
         _cons |   1.183848   .1413397     1.41   0.157     .9368515    1.495963
-------------+----------------------------------------------------------------
Ani_souhlas__ani_nesouhlas |
         2.sex |    .734074   .0515181    -4.40   0.000     .6397372     .842322
           age |   .9974305   .0020916    -1.23   0.220     .9933395    1.001538
         _cons |   1.257507   .1471583     1.96   0.050     .9997684    1.581689
-------------+----------------------------------------------------------------
Spíše_nesouhlasím |  (base outcome)
-------------+----------------------------------------------------------------
Rozhodně_nesouhlasím |
------------------------------------------------------------------------------
```

Obrázek 3 Ukázka multinomické logistické regrese

Multinomická logistická regrese se interpretuje pro každou kategorii závisle proměnné zvlášť ve vztahu k referenční kategorii. Na výše uvedeném obrázku proto například platí, že pokud je respondent ženského pohlaví, je o 26,6 % nižší pravděpodobnost (1-0,734), že bude „ani souhlasit, ani nesouhlasit“, ve srovnání s tím, že bude „spíše nesouhlasit“. Interpretace multinomického logistického modelu je proto extrémně složitá, a pokud je to jen trochu možné, použijeme raději ordinální logistickou regresi.

Podmínkou je ale stejná vzdálenost mezi jednotlivými variantami závisle proměnné, kterou nejlépe ověří Brantův test. Postup je takový, že nejprve provedeme ordinální logistickou regresi (příkaz OLOGIT) a následně zadáme příkaz BRANT, DETAIL. Ten zjistí, jestli je použití ordinální logistické regrese vhodné.

```
ologit gearn i.sex age
brant, detail
```

```
. brant, detail

Estimated coefficients from binary logits
```

Variable	y_gt_1	y_gt_2	y_gt_3	y_gt_4
sex				
2	0.250	0.303	0.415	0.410
	3.51	5.98	8.45	6.44
age	-0.016	-0.004	0.000	0.004
	-7.37	-2.60	0.31	1.99
_cons	2.602	0.696	-0.534	-1.905
	20.39	8.10	-6.46	-17.73

legend: b/t

Brant test of parallel regression assumption

	chi2	p>chi2	df
All	73.14	0.000	6
2.sex	7.92	0.048	3
age	65.00	0.000	3

A significant test statistic provides evidence that the parallel regression assumption has been violated.

Obrázek 4 Použití Brantova testu

V našem případě bohužel Brantův test doporučuje použít multinomický logistický model.

7.5 Kvalita regresních modelů

Při práci s regresními modely postupujeme vždy od jednodušších modelů ke složitějším. Sledujeme přitom, jak se mění kvalita modelu. Cílem je dosáhnout stavu, kdy je model co nejjednodušší, ale přitom co nejkvalitnější (protichůdné požadavky). Než se podíváme na příklad budování modelů, probereme jednotlivé ukazatele kvality.

U lineární regrese je hlavním ukazatelem kvality R^2, který vyjadřuje, do jaké míry model reprezentuje sociální realitu. Hodnota je sice udávána v desetinném čísle, ale po vynásobení stem dostáváme hodnotu v procentech (např. R^2 = 0,1234 znamená, že model vysvětluje realitu z 12,34 %). Samozřejmě platí, že čím vyšší je R^2, tím kvalitnější je model. U lineárních modelů je potřeba reportovat dvě hodnoty, kromě R^2 také počet případů N.

U logistické regrese ukazatel R^2 použít nemůžeme, i když ho Stata v podobě Pseudo R^2 nabízí. Ukazateli kvality logistického modelu jsou hodnoty likelihood ratio, AIC a BIC. Alespoň jeden z nich je potřeba – spolu s počtem případů N – reportovat spolu s modelem. Ani jeden z uvedených ukazatelů přitom neříká nic o absolutní kvalitě modelu (oproti R^2, které je použitelné i pro jeden samostatný model), slouží pouze ke srovnání kvality dvou modelů. Modely musí být do sebe vnořené (nested), to znamená, že jeden model musí obsahovat tytéž proměnné jako druhý model, případně nějaké navíc, a oba modely musí mít stejný počet případů N.

Poměr věrohodnosti (log-likelihood ratio, -2LL), obecně platí, že čím větší log-likelihood, tím lepší model. Akaikovo informační kriterium AIC podle některých zdrojů nevyžaduje, aby byly porovnávané modely do sebe vnořené, zohledňuje počet nezávislých proměnných. Platí, že čím menší AIC, tím lepší model. Bayesovo informační kritérium (BIC) penalizuje složité modely ještě silněji, lze ho ale teoreticky využít i pro modely s různým N. I v tomto případě volíme model, který má hodnotu BIC co nejnižší.

Charakteristiky modelu vypisují příkazy ESTAT IC nebo FITSTAT.

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	7098	-11269.74	-11233.26	6	22478.53	22519.73

Note: N=Obs used in calculating BIC; see [R] BIC note

Obrázek 5 Posuzování kvaliti regresního modelu

7.6 Postupné budování modelu

Při návrhu regresních modelů vycházíme, stejně jako při jiných sociologických analýzách, z teorie. Studiem literatury zjistíme, které nezávisle proměnné by měly ovlivňovat závisle proměnnou. Nevytváříme zvláštní regresní model pro

každou hypotézu, ale jeden model celkový, z kterého následně rozhodneme o platnosti všech hypotéz. Všechny modely prezentujeme v jedné přehledné tabulce včetně potřebných ukazatelů kvality modelu.

Příklad: Souvislost vzdělání a výše příjmu
Literatura: Teorie lidského kapitálu říká, že výše příjmu je ovlivněna především délkou praxe (většinou nahrazeno věkem) a stupněm dosaženého vzdělání (lidským kapitálem). Mincerova rovnice empiricky ukazuje, že je potřeba porovnávat přirozený logaritmus hrubého příjmu s nezávisle proměnnými věk, druhá mocnina věku, stupeň vzdělání. Nejprve si připravíme data:

```
gen prijem=payn
replace prijem=payn/12 if paynam==1 //pokud je příjem roční, převedeme ho na měsíční
replace prijem=prijem/(hours*4.5) //měsíční příjem převedeme na hodinový tak, že ho vydělíme počtem hodin odpracovaných týdně a počtem týdnů v měsíci
replace prijem=ln(prijem) //vypočteme logaritmus hodinového příjmu
gen age2=age*age //druhá mocnina věku
recode educ (0/4=1) (5/8=2) (9/11=3)
```

Nyní odhadneme první regresní model M1:

```
regress prijem i.educ age age2
```

```
. regress prijem i.educ age age2

      Source |       SS       df       MS              Number of obs =    3033
-------------+------------------------------           F(  4,  3028) =   43.26
       Model |  180.075668     4  45.0189171            Prob > F      =  0.0000
    Residual |  3151.16153  3028  1.04067422            R-squared     =  0.0541
-------------+------------------------------           Adj R-squared =  0.0528
       Total |  3331.2372   3032  1.09869301            Root MSE      =  1.0201

------------------------------------------------------------------------------
      prijem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |
          2  |   .3192163   .043651     7.31   0.000     .2336276    .4048049
          3  |   .6130378   .049264    12.44   0.000     .5164435    .7096321
             |
         age |   .0188083   .0096445    1.95   0.051    -.0001021    .0377186
        age2 |  -.0001265   .0001057   -1.20   0.231    -.0003337    .0000807
       _cons |   3.570125   .2155492   16.56   0.000     3.147488    3.992763
------------------------------------------------------------------------------
```

Obrázek 6 Regresní model M1

Vidíme, že regresní model vysvětluje realitu z 5,41 %. Dosažení maturitního vzdělání zvyšuje příjem o 32 %, dosažení vysokoškolského vzdělání o 61 % ve srovnání se vzděláním základním. Získání jednoho roku praxe zvýší příjem o 1,8 %. Předpokládáme ale, že na příjem má vliv i pohlaví respondenta, statistiky Eurostatu říkají, že Česko patří k zemím s nejvyšším genderovým rozdílem v příjmech. Přidáme proto do modelu M2 ještě pohlaví.

```
regress prijem i.educ age age2 i.sex
```

Kvalita modelu se mírně zvýšila, koeficienty se změnily (viz tabulka). Nyní nás zajímá, jak se liší návratnost vzdělání pro ženy v závislosti pro různé stupně vzdělání, tedy zda vysokoškolačky získají ze svého vzdělání stejně jako středoškolačky. Vytvoříme proto model M3 s přidanou interakcí mezi pohlavím a stupněm vzdělání.

```
regress prijem i.educ age age2 i.sex i.sex#i.educ
```

Všimněte si, že postupně jsme vytvořili čtyři hypotézy, ale výsledky budeme interpretovat z jediného regresního modelu:

H1: Čím více má člověk praxe, tím vyšší má příjem.
H2: Čím je člověk vzdělanější, tím vyšší má příjem.
H3: Ženy mají nižší příjem než muži bez ohledu na vzdělání i praxi.
H4: Absolvování vysoké školy dává ženám příležitost dohnat genderový rozdíl v příjmech.

Všechny modely shrneme do jediné tabulky, ze které pak výsledky slovně interpretujeme:

Tabulka 1 Determinanty výše přijmu

		M1	M2	M3
Vzdělání	Vyučen	Ref.	Ref.	Ref.
	Maturita	0,319***	0,330***	0,387***
	VŠ	0,613***	0,618***	0,683***
Věk		0,019*	0,024*	0,024*
Věk*Věk		-0,000	-0,000+	-0,000+
Pohlaví	Muž		Ref.	Ref.
	Žena		-0,211***	-0,131*
Interakce	Muž a Vyučen			Ref.
	Žena * Maturita			-0,112
	Žena * VŠ			-0,128
Konstanta		3,570***	3,558***	3,517***
N		3 033	3 033	3 033
R^2		0,0541	0,0641	0,0648

Statistická signifikance: *** p<0.001 ** p<0.01 * p<0.05 + p<0.1

Na základě modelu M3 nemůžeme vyvrátit hypotézu H1 (v našich datech skutečně s rostoucí praxí roste příjem) ani H2 (v našich datech rostoucí vzdělání skutečně zvyšuje příjem). Stejně tak hypotéza H3 musí zůstat podržena, v modelu M2 dokonce hodnota 21,1 % přibližně odpovídá rozdílu gender pay gap, který uvádí pro Českou republiku Eurostat (v modelu M3 bychom museli interpretovat vliv pohlaví spolu s vlivem interakce mezi pohlavím a vzděláním). Hypotézu H4 musíme zamítnout, VŠ žena má příjem nižší o (0,131 + 0,128) 25,9 %, zatímco žena s maturitou má příjem nižší jen o (0,131 + 0,112) 24,3 % nižší.

Nejčastější chyby:

Vytváření zvláštních modelů pro každou hypotézu – Je potřeba vytvořit „jeden velký model“, který kontroluje vliv ostatních proměnných.

Interpretace jen jednoho modelu – Je potřeba dokázat, že použitý model je ten nejlepší možný, proto je vhodné ukázat více modelů.

Chybějící charakteristiky modelu – U každého modelu musí být informace o jeho kvalitě. U lineární regrese N a R^2, u logistické regrese N a buď AIC nebo BIC.

Chybějící referenční hodnoty – U kategorizovaných proměnných je potřeba ukázat čtenáři, která hodnota je referenční.

Vyřazení nesignifikantních proměnných – Nesignifikance nezávisle proměnných ztěžuje jejich interpretaci, na druhou stranu i nesignifikantní proměnná slouží jako kontrolní proměnná, která očišťuje vliv ostatních proměnných. Pokud je její přítomnost v modelu zdůvodnitelná (z literatury nebo z logické úvahy), je lepší ji tam ponechat.