7. REGRESE

Regrese je statistická metoda, která umožňuje snadno odhalit vliv několika nezávisle proměnných (*independent variable*) na jednu závisle proměnnou (*dependent variable*). Předpokládá se, že vztah závisle a nezávisle proměnných je lineární (odtud lineární regrese) a rozložení obou proměnných je normální. Podle typu závisle proměnné vybíráme vhodný typ regrese:

Kardinální:

- lineární regrese, ve Statě příkaz REGRESS, pokud jsou data intervalová (nemají neutrální hodnotu, tj. chybí nula)
- Poissonova regrese, ve Statě příkaz POISSON pokud jsou data poměrová (mají rozsah od nuly do nekonečna, např. počet pokusů o sebevraždu)

Ordinální a nominální:

- binární logistická regrese (hodnoty 0,1), ve Statě příkaz LOGIT
- ordinální logistická regrese (uspořádané hodnoty, vzdálenosti mezi jednotlivými hodnotami jsou stejné), ve Statě příkaz OLOGIT
- multinomická logistická regrese (hodnoty nemusí být uspořádány a mezi nimi nemusí být stejná vzdálenost), ve Statě příkaz MLOGIT.

7.1 Lineární regrese

Odhadem koeficientů regresního modelu hledáme rovnici přímky, která ideálním způsobem proloží body jednotlivých pozorování. Základní rovnice přímky má podobu:

$$y = a + bx + e$$

kde y je závisle proměnná, a je konstanta vypočtená z regresního modelu (průsečík přímky s osou y pro x=0), b je koeficient regresního modelu vypočtený z regresního modelu (sklon přímky proti ose x), x je hodnota nezávisle proměnné, e jsou nevysvětlení rezidua.

Přidáváním dalších nezávisle proměnných x_1, x_2, x_3… se rovnice komplikuje následovně:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + … + e$$

REGRESS – odhadne koeficienty regresního modelu zadaného za příkazem REGRESS. Na první místo se píše závisle proměnná (vysvětlovaná), za ni se v libovolném pořadí zapisují nezávisle proměnné (vysvětlované, prediktory, determinanty)

```
regress payn
regress payn age
```

- R-squared: z kolika procent model vysvětluje sociální realitu. Po vynásobení stovkou dostaneme míru vysvětlování v procentech.
- Prob > F: statistická významnost celého modelu (platí obvyklá hranice 0.05)
- P>|t|: statistická významnost jednotlivých vysvětlujících proměnných (platí obvyklá hranice 0.05)

Kategorizované proměnné je potřeba zadat jako tzv. dummy proměnné, což v praxi znamená, že pro každou hodnotu proměnné je vytvořena nová proměnná nabývající hodnot 0 a 1. Např. pro proměnnou dny_v_tydnu by bylo vytvořeno sedm dummy proměnných dny_v_tydnu_pondeli (nabývající hodnoty 1 v pondělí a hodnot 0 v jiných dnech), dny_v_tydnu_utery (nabývající hodnoty 1 v úterý a hodnot 0 v jiných dnech) apod. Při odhadu regresních modelů stačí jednoduše před každou kategorizovanou (tj. ordinální nebo nominální) zapsat písmeno i s tečkou. Stata pak zvolí první hodnotu (tedy např. pondělí) jako tzv. referenční a vysvětluje efekt následujících hodnot ve srovnání s touto referenční hodnotou.

```
regress payn age i.educ
```

Pokud nás zajímá souvislost dvou či více proměnných, necháme odhadnout tzv. interakci. Stata pak vypočte, jak podobu výsledné přímky ovlivňuje kombinace dvou nezávisle proměnných (např. kombinace věku a pohlaví). Interakce se zadává jednoduše pomocí znaku # mezi dvěma proměnnými.

```
regress payn age i.educ i.sex i.sex#i.educ
```

Jelikož věk není vhodné pojímat jako kategorizovanou proměnnou (Stata vypočte interakci pro každou hodnotu věku a každé pohlaví, např. tedy 17 let pro muže, 17 let pro ženy, 18 let pro muže, 18 let pro ženy …). Pomocí znaku c a tečka můžeme Statě nařídit, aby danou proměnnou považovala za spojitou.

```
regress payn age i.educ i.sex c.age#i.sex
```

```
. regress payn age i.educ i.sex c.age#i.sex

      Source |       SS       df       MS              Number of obs =    3229
-------------+------------------------------           F(  5,  3223) =   29.48
       Model |  3321752.48     5  664350.496            Prob > F      =  0.0000
    Residual |  72632909.2  3223  22535.808            R-squared     =  0.0437
-------------+------------------------------           Adj R-squared =  0.0422
       Total |  75954661.7  3228  23529.9447           Root MSE      =  150.12

------------------------------------------------------------------------------
        payn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.444558   .2961792     4.88   0.000     .8638396    2.025277
             |
        educ |
          2  |    29.12699   6.278198     4.64   0.000     16.81733    41.43666
          3  |     63.8274   6.879852     9.28   0.000     50.33807    77.31673
             |
       2.sex |   -32.33559   19.99666    -1.62   0.106    -71.54306    6.871874
             |
   sex#c.age |
          2  |   .2414819   .4338899     0.56   0.578    -.6092461     1.09221
             |
       _cons |    25.97365   14.33949     1.81   0.070    -2.141781    54.08909
------------------------------------------------------------------------------
```

Obrázek 1 Ukázka práce s příkazem regress

PREDICT – vypočte hodnoty proměnné podle posledního vypočteného modelu. Důležitými parametry jsou XB, který odhadne hodnoty lineárního modelu, a RES, který vypočte hodnoty reziduálů (rozdíl mezi naměřenou a vypočtenou hodnotou)

```
predict novapromenna, xb res
```

MARGINS – odhadne hodnoty závisle proměnné při udržení ostatních proměnných na konstantní hladině. Pomocí příkazu MARGINSPLOT je pak možno vykreslit průběh proměnné. Následující příkazy odhadnou „marginové hodnoty" proměnné payn pro věk v rozsahu 0-100 let tabelovaného po 5 letech (všimněte si, že v původním souboru má věk rozsah pouze 0-94, to příkazu margine nebrání vypočíst odhady i pro věk mimo tento rozsah) odděleně pro obě pohlaví.

```
margins, at(age=(0(5)100)) over(sex)
marginsplot
```