

## Panelové dáta v programe Gretl

Martin Lukáčik, Adriana Lukáčiková, Karol Szomolányi

Panelové dáta sú kombinované prierezové a časové údaje. Pri panelových údajoch existuje časový rad pre každú entitu použitú v rámci prierezového výberu. Najčastejšie sa panelové dáta využívajú na skúmanie časového vývoja rôznych jednotiek z toho istého sektora, trhu alebo geografického celku, pričom typické sú rozsiahla prierezová štruktúra a len niekoľko časových období.

Panelové dáta sú z hľadiska zvládnutia štatistických metód náročnejšou oblasťou ako samostatné prierezové dáta a časové rady, preto nebudeme prezentovať všetky možnosti ich analýzy, ale zameriame sa iba na základy. Sem rátame najmä pochopenie postupu vloženia panelových údajov do programu Gretl a pochopenie postupu ako sa odhadujú modely s fixnými a náhodnými efektmi. Náročnejšie postupy by následne už nemali byť pre čitateľa problémom práce so softvérom, ale iba problémom zvládnutia metodológie.

### Úvod

Za základný regresný model panelových dát považuje Greene (2003) model:

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \dots + \beta_k x_{itk} + \alpha_1 z_{i1} + \alpha_2 z_{i2} + \dots + \alpha_q z_{iq} + u_{it},$$

kde index  $i$  označuje prierezový rozmer  $i = 1, \dots, n$ , index  $t$  časový rozmer  $t = 1, \dots, T$ , premenné  $X_1$  až  $X_k$  sú vysvetľujúce premenné nezahŕňajúce vektor jednotiek a premenné  $Z_1$  až  $Z_q$  predstavujú individuálne efekty – rôznorodosť, ktorou sa môže odlišovať jednotlivec alebo celá skupina od ostatných entít – sem sa zaraďuje prípadný vektor jednotiek. Individuálne efekty sa nemenia sa s časom.

Na základe uvedeného rámca rozlíšime a ukážeme ako sa odhadujú tri prípady:

- **spojený regresný model** (Pooled Regression) – ak individuálnym efektom je iba vektor jednotiek, čo znamená, že jediný parameter  $\alpha$  je spoločnou konštantou:

$$y_{it} = \alpha + \beta_1 x_{it1} + \beta_2 x_{it2} + \dots + \beta_k x_{itk} + u_{it},$$

- **model s fixnými efektmi** (Fixed Effects Model – FEM) – ak individuálne efekty  $Z_1$  až  $Z_q$  sú nepozorovateľné, ale korelované s vysvetľujúcimi premennými, tak riešením je zahrnúť všetky efekty do odhadnuteľného podmieneného priemeru pomocou vzťahu  $\alpha_i = \alpha_1 z_{i1} + \alpha_2 z_{i2} + \dots + \alpha_q z_{iq}$  a model FEM má tvar:

$$y_{it} = \alpha_i + \beta_1 x_{it1} + \beta_2 x_{it2} + \dots + \beta_k x_{itk} + u_{it},$$

– fixný efekt  $\alpha_i$  znamená špecifickú konštantu pre každú prierezovú jednotku,

- **model s náhodnými efektmi** (Random Effects Model – *REM*) – ak individuálne efekty  $Z_1$  až  $Z_q$  sú nepozorovateľné, ale nekorelované s vysvetľujúcimi premennými, tak riešením je zložená náhodná zložka  $\varepsilon_i + u_{it}$ , ktorá okrem pôvodnej predpokladá aj špecifickú náhodnú zložku pre každú prierezovú jednotku a model *REM* má tvar:

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \dots + \beta_k x_{itk} + (\alpha + \varepsilon_i) + u_{it}.$$

### Načítanie panelových dát do programu Gretl

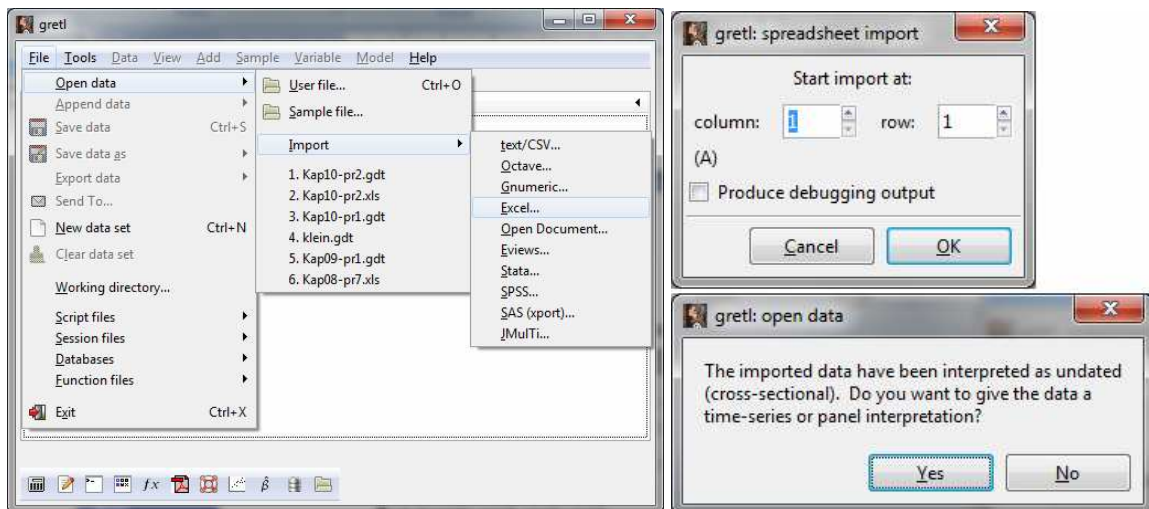
Najrýchlejším spôsobom zavádzania údajov tvoriacich panelové dáta do programu Gretl je ich vhodná príprava v tabuľkovom procesore (napríklad Microsoft Excel). Za vhodnú prípravu môžeme považovať tabuľky údajov naskladaných za sebou, pričom údaje môžu byť naskladané prierezo-vo alebo časovo. Ako príklad údajov „naskladaných časovo“ (po rade jednej prierezovej jednotky nasleduje rad druhej jednotky, atď.) môžeme uviesť tabuľku s údajmi o leteckých spoločnostiach z učebnice Lukáčiková a Lukáčik (2008) na obrázku 1, pričom pokračovanie pôvodných stĺpcov ako ďalšie stĺpce (ako to bolo zobrazené v učebnici) samozrejme nie je dovolené. Údaje obsahujú celkové náklady TC, produkciu Q, cenu benzínu P a koeficient vyťaženia lietadiel KV 6 firiem za 15 rokov.

Obrázok 1: Usporiadanie dát –naskladané časové rady

	A	B	C	D	E	F
1	$i$	$t$	TC	Q	P	KV
2	1	1	1140640	0.952757	106650	0.534487
3	1	2	1215690	0.986757	110307	0.532328
4	1	3	1309570	1.09198	110574	0.547736
5	1	4	1511530	1.17578	121974	0.540846
6	1	5	1676730	1.16017	196606	0.591167
7	1	6	1823740	1.17376	265609	0.575417
8	1	7	2022890	1.29051	263451	0.594495
9	1	8	2314760	1.39067	316411	0.597409
10	1	9	2639160	1.61273	384110	0.638522
11	1	10	3247620	1.82544	569251	0.676287
12	1	11	3787750	1.54604	871636	0.605735
13	1	12	3867750	1.5279	997239	0.61436
14	1	13	3996020	1.6602	938002	0.633366
15	1	14	4282880	1.82231	859572	0.650117
16	1	15	4748320	1.93646	823411	0.625603
17	2	1	569292	0.520635	103795	0.490851
18	2	2	640614	0.534627	111477	0.473449
19	2	3	777655	0.655192	118664	0.503013
20	2	4	999294	0.791575	114797	0.512501
21	2	5	1203970	0.842945	215322	0.566782
22	2	6	1358100	0.852892	281704	0.558133
23	2	7	1501350	0.922843	304818	0.558799
24	2	8	1709270	1	348609	0.57207
25	2	9	2025400	1.19845	374579	0.624763
26	2	10	2548370	1.34067	544109	0.628706
27	2	11	3137740	1.32624	853356	0.58915
28	2	12	3557700	1.24852	1003200	0.532612
29	2	13	3717740	1.25432	941977	0.526652
30	2	14	3962370	1.37177	856533	0.540163
31	2	15	4209390	1.38974	821361	0.528775
32	3	1	286298	0.262424	118788	0.524334
33	3	2	309290	0.266433	123798	0.537185

V programe Gretl pomocou ponuky *File/Open\_data/Import/Excel* načítame pripravené údaje z tabuľkového procesora, pričom sa na konci importu ponúkne zmena interpretácie dát.

Obrázok 2: Postup načítania údajov z pripraveného súboru v tabuľkovom procesore



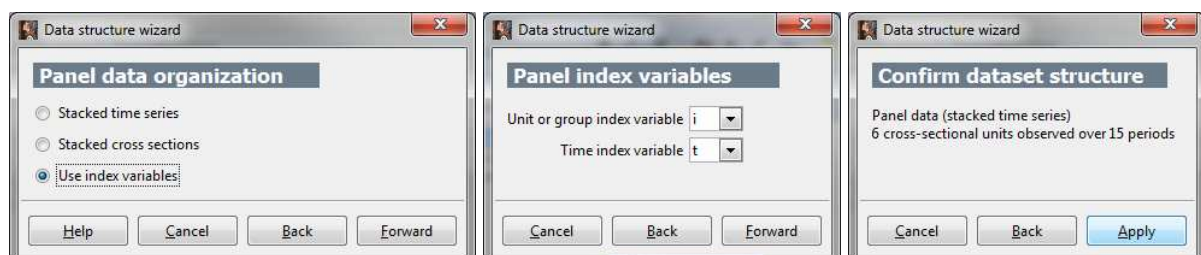
Po načítaní súboru zmeníme štruktúru údajov voľbou *Yes*, pričom zvolíme štruktúru údajov typu *Panel* a organizáciu dát vo forme *Stacked time series*. Posledným krokom sprievodcu zmenou dátovej štruktúry je spresnenie počtu prierezových jednotiek (v našom prípade 6) alebo časových období, pričom druhý z údajov je vždy dopyčovaný.

Obrázok 3: Spresnenie typu štruktúry panelových údajov naskladaných časovo



Organizáciu dát môžeme v tomto prípade riešiť aj pomocou indexových premenných, teda radov *i* a *t*, ktoré použijeme ako identifikátory pre prierezové jednotky *Unit or group index variables* a pre časový index *Time index variable*. Na konci sa zobrazí oznam o počte prierezových jednotiek a časových období v danom súbore.

Obrázok 4: Pracovný zošit programu EViews s panelovými dátami



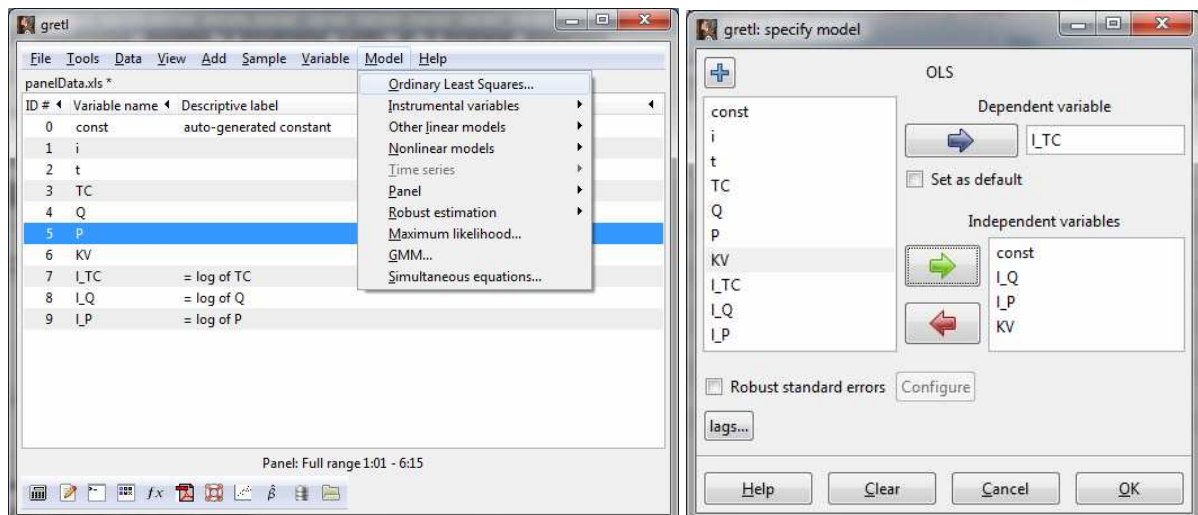
## Analýza panelových dát v programe Gretl

Najjednoduchší prípad, ktorým je **spojený regresný model (pool)**, predstavuje naivný prístup, v ktorom sa predpokladá, že absolútny člen aj všetky parametre pri vysvetľujúcich premenných sú pre všetky prierezové jednotky rovnaké. Spojený model má všeobecný tvar:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \alpha + \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \alpha + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

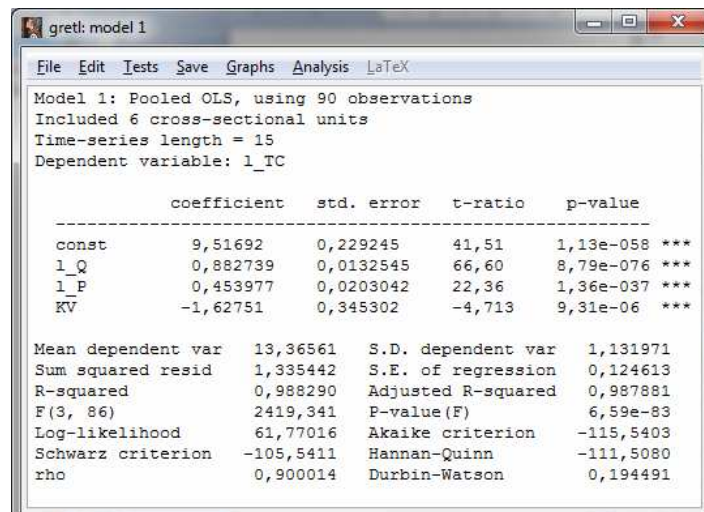
Odhad spojeného regresného modelu v programe Gretl je v prípade štruktúry údajov typu panelové dáta analógiou obyčajného odhadu lineárneho modelu v iných typoch údajov pomocou ponuky *Model/Ordinary\_Least\_Squares*.

Obrázok 5: Odhad parametrov spojeného regresného modelu



Po voľbe závislej a nezávislých premenných získame odhad spojeného modelu (*pool*).

Obrázok 6: Výsledný spojený regresný model



```
Model 1: Pooled OLS, using 90 observations
Included 6 cross-sectional units
Time-series length = 15
Dependent variable: l_TC
```

	coefficient	std. error	t-ratio	p-value
const	9,51692	0,229245	41,51	1,13e-058 ***
l_Q	0,882739	0,0132545	66,60	8,79e-076 ***
l_P	0,453977	0,0203042	22,36	1,36e-037 ***
KV	-1,62751	0,345302	-4,713	9,31e-06 ***

Mean dependent var	13,36561	S.D. dependent var	1,131971
Sum squared resid	1,335442	S.E. of regression	0,124613
R-squared	0,988290	Adjusted R-squared	0,987881
F(3, 86)	2419,341	P-value(F)	6,59e-83
Log-likelihood	61,77016	Akaike criterion	-115,5403
Schwarz criterion	-105,5411	Hannan-Quinn	-111,5080
rho	0,900014	Durbin-Watson	0,194491

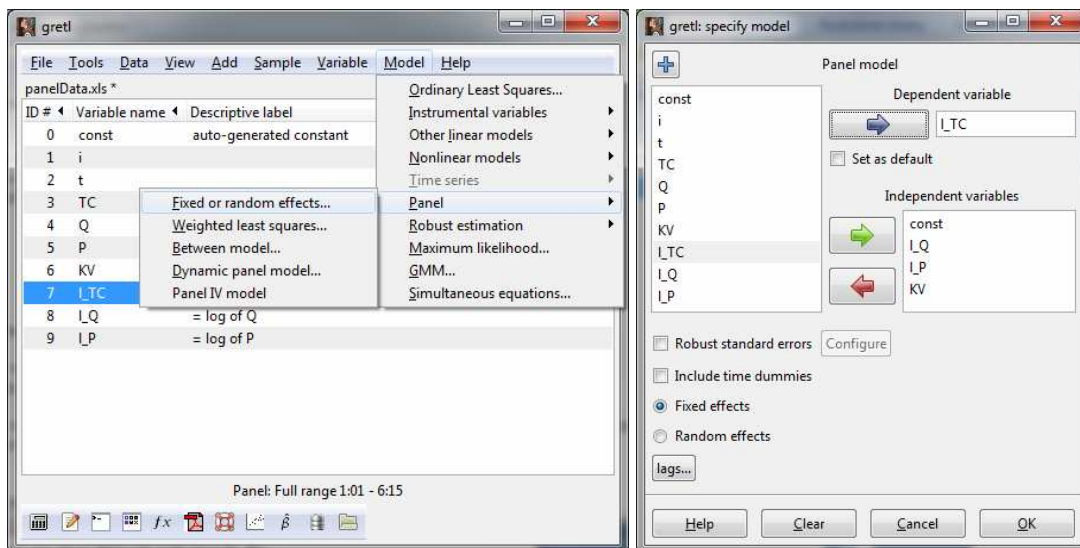


**Model s fixnými efektmi (FEM)** na rozdiel od spojeného regresného modelu predpokladá rôznorodosť prierezových jednotiek v absolútnych členoch. Model FEM má tvar:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

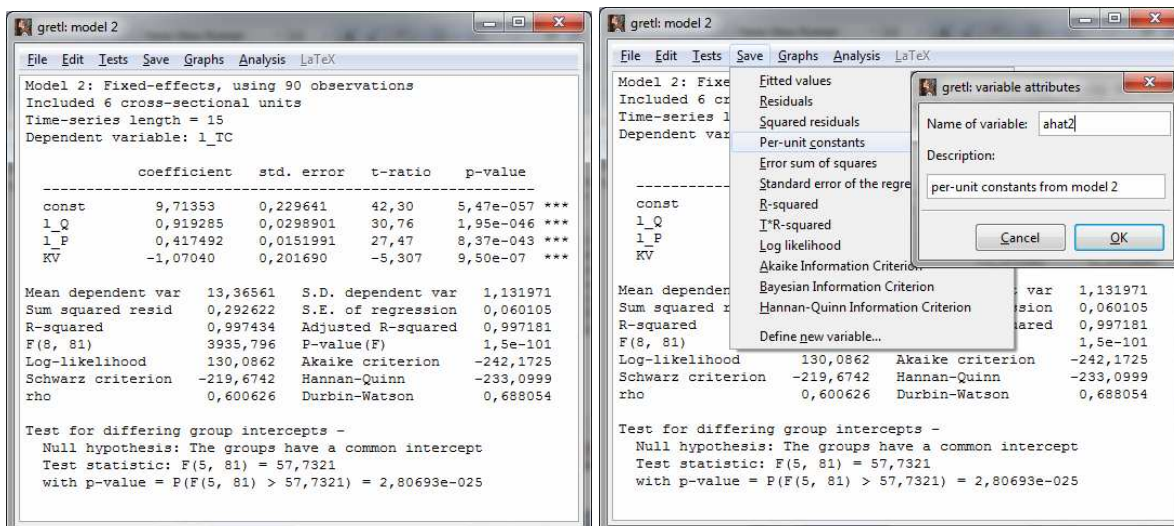
Vidíme, že v modeli predstavujú stĺpce matice  $\mathbf{D}$  umelé premenné  $D_1$  až  $D_n$ , ktoré nadobúdajú hodnotu  $d_{it} = 1$  pre  $i$ -tú prierezovú jednotku, a hodnotu  $d_{it} = 0$  pre všetky ostatné prierezové jednotky. Odhad v programe Gretl spustíme voľbou *Model/Panel/Fixed\_or\_random\_effects* a zaškrtnutím voľby *Fixed effects* pri špecifikácii modelu (na obrázku 7).

Obrázok 7: Odhad parametrov modelu FEM



Výsledkom je odhadnutý model FEM na obrázku 8. Efekty jednotlivých prierezových jednotiek sa dajú získať ako nová premenná **ahat2** cez ponuku *Save/Per-unit\_constants*.

Obrázok 8: Výsledný model FEM a postup pre získanie výpisu fixných efektov



Testovanie významnosti parametrov  $\alpha_i$  obyčajne nemá veľký význam. Preto Gretl túto hodnotu ani neposkytuje. Rozdielnosť medzi jednotlivými prierezovými jednotkami sa testuje pomocou  $F$ -testu porovnávajúceho model  $FEM$  a model  $pool$ . Testovacia štatistika zobrazená pod odhadom rovnice ako *Test for differing group intercepts* má tvar:

$$F = \frac{(RSS_{pool} - RSS_{FEM}) / (n-1)}{RSS_{FEM} / (nT - k - n)}$$

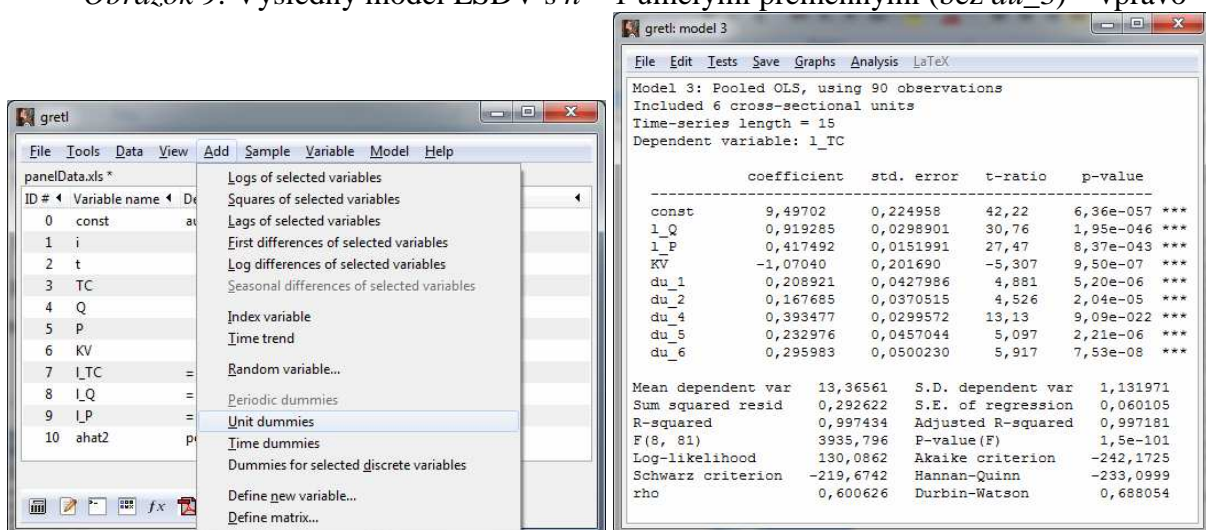
a porovnávame ju s tabuľkovou hodnotou  $F(n-1, nT - k - n)$  na zvolenej hladine významnosti  $\alpha$ . Ak je hodnota štatistiky väčšia ako tabuľková hodnota, zamietame nulovú hypotézu, že prierezové jednotky majú rovnaké absolútne členy.

Kvôli umelým premenným sa tento *model* zvykne nazývať aj **LSDV** (Least Squares Dummy Variable) a takto ho môžeme po vytvorení umelých premenných pre každú prierezovú jednotku odhadnúť. V programe Gretl sa vytvorí v ponuke *Add/Unit\_dummies* – obrázok 9 vľavo. Model môžeme odhadnúť ako regresný model bez konštanty, alebo sa zvolí jedna prierezová jednotka za základnú skupinu (napríklad tretia), ktorej hodnotu v modeli LSDV bude predstavovať absolútny člen a využijeme len  $n - 1$  umelých premenných:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \alpha_1 + \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{i} & \cdots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \begin{bmatrix} \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_1 \\ \vdots \\ \alpha_n - \alpha_1 \end{bmatrix} + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \alpha_1 + \mathbf{D}_1 \boldsymbol{\alpha}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

V modeli matica  $\mathbf{D}_1$  predstavuje maticu  $\mathbf{D}$  bez prvého stĺpca a vektor  $\boldsymbol{\alpha}^*$  je  $n-1$  prvkový vektor diferencujúcich absolútnych členov vzťahujúcich sa na absolútny člen základnej skupiny.

Obrázok 9: Výsledný model LSDV s  $n - 1$  umelými premennými (bez  $du\_3$ ) – vpravo



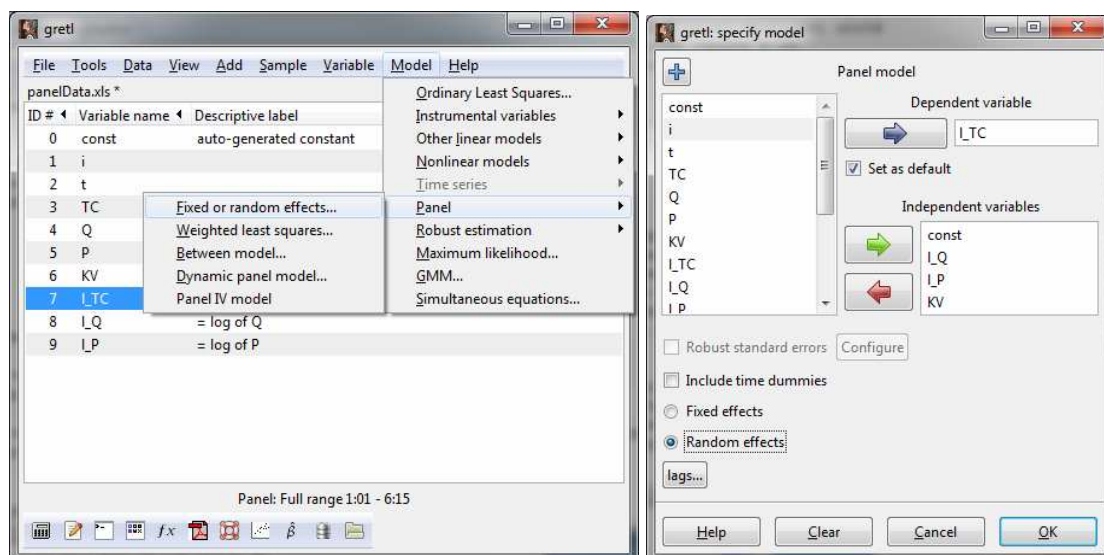
V modeli s fixnými efektmi sa odlišnosť medzi priereznými jednotkami uvažovala ako posun regresnej funkcie. Ak individuálne vplyvy pôsobiace na prierezné jednotky nie sú korelované s vysvetľujúcimi premennými celého panelu, tak by bolo vhodnejšie modelovať jednotlivé absolútne členy pre prierezné údaje ako náhodne rozdelené.

**Model s náhodnými efektmi (REM)** má tvar:

$$y_{it} = \alpha + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + \varepsilon_i + u_{it} = \alpha + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + v_{it},$$

kde spojením náhodnej zložky konkrétneho pozorovania v prierezovej jednotke  $u_{it}$  a náhodnej zložky špecifickej pre priereznú jednotku  $\varepsilon_i$  dostávame zloženú náhodnú zložku  $v_{it}$ . Absolútny člen  $\alpha$  predstavuje v modeli priemer prierezných absolútnych členov a náhodná zložka špecifická pre priereznú jednotku je náhodnou odchýlkou od tohto priemeru. Pri odhade v programe Gretl cez ponuku *Model/Panel/Fixed\_or\_random\_effects* zaškrtneme voľbu *Random effects* (na obrázku 10).

Obrázok 10: Odhad parametrov modelu REM



Výsledkom je odhadnutý model REM na obrázku 11. Na voľbu medzi modelmi *FEM* a *REM* sa využíva **Hausmanov test** špecifickácie, ktorého testovacia štatistika je zobrazená pod odhadom modelu na konci výpisu. Nulová hypotéza predpokladá, že odhady parametrov zovšeobecnenej metódy najmenších štvorcov v modeli REM a metódy najmenších štvorcov v modeli FEM sú konzistentné, a tým odhad metódou najmenších štvorcov nie je efektívny. V alternatívnej hypotéze je len metóda najmenších štvorcov konzistentná. Ak je hodnota štatistiky  $H > \chi^2_c$ , tak môžeme zamietnuť nulovú hypotézu o konzistentnosti oboch estimátorov a vhodnejší je model *FEM*. Ak je hodnota štatistiky  $H < \chi^2_c$ , tak nemôžeme zamietnuť nulovú hypotézu a odporúčaným bude model *REM*.

Obrázok 11: Výsledný odhad modelu REM s Hausmanovým testom

```

gretl: model 4
File Edit Tests Save Graphs Analysis LaTeX
Model 4: Random-effects (GLS), using 90 observations
Included 6 cross-sectional units
Time-series length = 15
Dependent variable: l_TC

      coefficient   std. error   t-ratio   p-value
-----
const      9,62791      0,210164    45,81     3,40e-062 ***
l_Q        0,906681      0,0256249   35,38     4,95e-053 ***
l_P        0,422778      0,0140248   30,15     1,72e-047 ***
KV         -1,06450      0,200070    -5,321    8,10e-07 ***

Mean dependent var   13,36561   S.D. dependent var   1,131971
Sum squared resid    1,495397   S.E. of regression    0,131105
Log-likelihood        56,67935   Akaike criterion     -105,3587
Schwarz criterion    -95,35945   Hannan-Quinn         -101,3264

'Within' variance = 0,00361262
'Between' variance = 0,0158381
theta used for quasi-demeaning = 0,876685

Breusch-Pagan test -
Null hypothesis: Variance of the unit-specific error = 0
Asymptotic test statistic: Chi-square(1) = 334,85
with p-value = 8,44102e-075

Hausman test -
Null hypothesis: GLS estimates are consistent
Asymptotic test statistic: Chi-square(3) = 3,52343
with p-value = 0,317735
    
```

Ak nechceme odhadovať všetky modely, ale stačí nám komplexná správa, tak priamo v okne spojeného modelu *pool* sa dá cez *Tests/Panel\_diagnostics* získať výpis na obrázku 12.

Obrázok 12: Diagnostická komplexná správa (na obrázku rozdelená kvôli rozsiahlosti)

```

gretl: panel model diagnostics
Diagnostics: assuming a balanced panel with 6 cross-sectional units
observed over 15 periods

Fixed effects estimator
allows for differing intercepts by cross-sectional unit
slope standard errors in parentheses, p-values in brackets

const:      9,7135      (0,22964)      [0,00000]
l_Q:        0,91928      (0,02985)      [0,00000]
l_P:        0,41749      (0,015195)     [0,00000]
KV:         -1,0704      (0,20169)      [0,00000]

6 group means were subtracted from the data

Residual variance: 0,292622/(90 - 9) = 0,00361262
Joint significance of differing group means:
F(5, 81) = 57,7321 with p-value 2,80693e-025
(A low p-value counts against the null hypothesis that the pooled OLS model
is adequate, in favor of the fixed effects alternative.)

Means of pooled OLS residuals for cross-sectional units:

unit 1:      0,068869
unit 2:     -0,013878
unit 3:     -0,19422
unit 4:      0,15273
unit 5:     -0,021583
unit 6:      0,0080906
    
```

```

gretl: panel model diagnostics
Breusch-Pagan test statistic:
LM = 334,85 with p-value = prob(chi-square(1) > 334,85) = 8,44102e-075
(A low p-value counts against the null hypothesis that the pooled OLS model
is adequate, in favor of the random effects alternative.)

Variance estimators:
between = 0,0158381
within = 0,00361262
theta used for quasi-demeaning = 0,876685

Random effects estimator
allows for a unit-specific component to the error term
(standard errors in parentheses, p-values in brackets)

const:      9,6279      (0,21016)      [0,00000]
l_Q:        0,90668      (0,025625)     [0,00000]
l_P:        0,42278      (0,014025)     [0,00000]
KV:         -1,0645      (0,20007)      [0,00000]

Hausman test statistic:
H = 3,52343 with p-value = prob(chi-square(3) > 3,52343) = 0,317735
(A low p-value counts against the null hypothesis that the random effects
model is consistent, in favor of the fixed effects model.)
    
```

### Literatúra:

- [1] GREENE, W. H.: *Econometric Analysis*, 4. vyd. New Jersey: Prentice-Hall, 2003.
- [2] LUKÁČIKOVÁ, A. – LUKÁČIK, M.: *Ekonometrické modelovanie s aplikáciami*. Bratislava: EKONÓM, 2008.
- [3] COTTRELL, A.: *Gretl User's Guide*. Gnu Regression, Econometrics and Time-series, 2011.
- [4] Internetové stránky k programu Gretl.