GMMCOVEARN: A Stata Module for GMM Estimation of the Covariance Structure of Earnings.

Aedín Doris*, Donal O'Neill & Olive Sweetman***

Abstract

This note describes **gmmcovearn** *a user-written Stata package that performs GMM estimation of the covariance structure of earnings for a variety of models. The program decomposes the variance of earnings into a permanent and transitory component using the GMM estimator. The program incorporates both time factor loadings and cohort factor loadings on both the transitory and permanent component, allows the transitory component to follow either an AR or an ARMA process and allows for random heterogeneous growth in the permanent component. The program is used in recent papers by Doris et al (2010a, 2010b).*

JEL Codes: J31, D31

Keywords: Stata, Permanent Inequality, Transitory Inequality, Generalized Method of Moments, Covariance Structure of Earnings

---

[**]Corresponding author: National University of Ireland Maynooth and IZA, Bonn and corresponding author. E-mail: donal.oneill@nuim.ie, (tel 353-1-7083555: fax 353-1-7083934; address Rhetoric House, NUI Maynooth, Maynooth, Co. Kildare, Ireland.
[*]National University of Ireland, Maynooth.

1. Introduction

In recent years there has been a rapid growth in the number of studies that have used the Generalised Method of Moments (GMM) estimator to estimate the covariance structure of earnings (e.g. Moffitt and Gottschalk (1995, 2002, 2008), Dickens (2000), Haider (2001), Ramos (2003), Baker and Solon (2003), Capellari (2004), Gustavsson (2004), Daly and Valetta (2007) Doris et al (2010a, 2010b)). In this note we describe **gmmcovearn** a user-written Stata program that simplified the estimation of these models.

2. GMM Approach to Estimating Permanent and Transitory Inequality

Following others we write earnings, y_{it}, as the sum of a permanent component, α_i, due for example to fixed characteristics such as the level of education, and a transitory one, v_{it}, reflecting temporary shocks that affect the individual or the labour market. That is

$$y_{bit} = c_{pb} p_t (\alpha_i + \beta_i X_{it}) + c_{tb} \lambda_t v_{it} \qquad (2)$$

where α_i and β_i are random variables with mean zero and variances σ_α^2 and σ_β^2 and covariance respectively $\sigma_{\alpha\beta}$. X_{it} is the age or experience of person i at time t. Only α_i is present in models without heterogeneous growth. p_t and λ_t are ‘factor loadings’ that allow these variances to change over time in a way that is common across individuals, while the c_{pb} and the c_{tb} are terms that allow the components to differ by cohorts. Our objective is to identify the separate roles played by the permanent and transitory shocks in determining inequality.

Persistence in the transitory shocks, v_{it}, is modelled using either an AR(1) or ARMA(1,1) process, with AR parameter ρ and MA parameter θ.

$$v_{it} = \rho v_{it-1} + \theta\varepsilon_{it-1} + \varepsilon_{it}$$

where ε_{it} is random variable with variance σ_ε^2 and the variance of v_1 is given by σ_{v1}^2.

The model is estimated by GMM using the Identity weighting matrix, whereby sample moments are matched to population moments. In this specification, the true variance-covariance matrix for cohort b has diagonal elements:

$$\sigma_{b1}^2 = c_{pb}^2 p_1^2(\sigma_\alpha^2 + \sigma_\beta^2 \overline{X_{b1}^2} + 2\sigma_{\alpha\beta}\overline{X_{b1}}) + c_{tb}^2\lambda_1^2\sigma_{v1}^2, \text{ for } t = 1$$

(3)

$$\sigma_{bt}^2 = c_{pb}^2 p_t^2(\sigma_\alpha^2 + \sigma_\beta^2 \overline{X_{b1}^2} + 2\sigma_{\alpha\beta}\overline{X_{b1}}) + c_{tb}^2\lambda_t^2(\rho^{2t-2}\sigma_{v1}^2 + K\sum_{w=0}^{t-2}\rho^{2w}), \text{ for } t > 1$$

and off-diagonal elements:

$$Cov_b(y_t, y_{t+s}) = c_{pb}^2 p_t\, p_{t+s}(\sigma_\alpha^2 + \sigma_\beta^2\overline{X_{bt}X_{bt+s}} + \sigma_{\alpha\beta}(\overline{X_{bt}} + \overline{X_{bt+s}})) + c_{tb}^2\lambda_t\, \lambda_{t+s}(\rho^s\sigma_{v1}^2 + \rho^{s-1}\theta\sigma_\varepsilon^2)$$
$$\text{, for } t = 1, s > 0$$

(4)

$$Cov_b(y_t, y_{t+s}) = c_{pb}^2 p_t\, p_{t+s}(\sigma_\alpha^2 + \sigma_\beta^2\overline{X_{bt}X_{bt+s}} + \sigma_{\alpha\beta}(\overline{X_{bt}} + \overline{X_{bt+s}})) + c_{tb}^2\lambda_t\, \lambda_{t+s}(\rho^{2t+s-2}\sigma_{v1}^2 + \rho^s K\sum_{w=0}^{t-2}\rho^{2w} + \rho^{s-1}\theta\sigma_\varepsilon^2)$$
$$\text{, for } t > 1, s > 0$$

where $K = \sigma_\varepsilon^2(1 + \theta^2 + 2\rho\theta)$, $\overline{X_{bt}}$ is the average age of cohort *b* at time *t*, and $\overline{X_{bt}^2}$ is the average value of age-squared for cohort *b* at time *t.*

The parameter vector to be estimated is given by $\varphi = \{\sigma_\alpha^2, \rho, \sigma_\varepsilon^2, \sigma_{v1}^2, p_1 ... p_T, \lambda_1 ... \lambda_T, c_{p1} .. c_{pC}, c_{t1} .. c_{tC}, \sigma_\beta^2, \sigma_{\alpha\beta}, \theta\}$. Identification requires a normalization of the factor loadings; in keeping with the literature, we set $\lambda_{1,}$ p_1, c_{p1} and c_{t1} equal to one. We then use this parameter vector to recover the individual components of aggregate inequality.

3. The gmmcovearn command

3.1 Syntax

gmmcovearn earningsvar, yearn(#) modeln(#) cohortn(#) agevar(#) firstyr(#) newdatname(string) stvalue() agevar(string) cohortvar(string) firstcohort(#)

3.2 Required Options

modeln(#) specifies the type of model to be estimated– default is AR, no-cohorts, no heterogeneity - model(1)}

ARMA, no-cohorts, no heterogeneity - model(2)
AR, no-cohorts, heterogeneity - model(3)
ARMA, no-cohorts, heterogeneity - model(4)
AR, cohorts, no heterogeneity - model(5)
ARMA, cohorts, no heterogeneity - model(6)
AR, cohorts, heterogeneity - model(7)
ARMA, cohorts, heterogeneity - model(8)

Yearn(#)specifies the number of years used for analysis.

3.3 Other options

Stvalue() specifies the starting values for the estimation. Values are entered in the following order, separated by commas: sigalpha, rho, sigv1, sige, l2-lT, p2-pT, cp2-cpC, ct2-ctC, sigbeta, covalphabeta, theta; which starting values are included depends on the model chosen.

Cohortn(#) specifies the number of cohorts used for analysis – default is **cohortn(1)**

Newdataname(string) specifies the name of the new file containing the moments and if a heterogeneous model is specified average age and average(age^2)

Graph(1) 1 if the user wants the a graphical display of the estimated variance decomposition - default is **graph(0)**

Agevar(string) specifies the name of the age variable to be used for heteroegenous growth models default is **agevar(age)**

Firstyr(#) specifies the numeric label attached to the first wave of earnings – default is **firstyr(1)**

Cohortvar(#) specifies the name of the cohort indicator- default is **cohortvar(cohort)**

Firstcohort(#) specifies the numeric indicator of the first cohort. Cohorts are then assumed to be labeled from **firstcohort** to **firstcohort** + **cohortn** -1.

3.4 Description:

gmmcovearn estimates parameters of the covariance structure of earnings using the earnings variable specified in *earningsvar*. The command makes use of an additional programs **nlcovearndfinalv1** that must be downloaded along with gmmcovearn. A detailed analysis of this approach to estimating the covariance structure of earnings can be found in Doris et al (2010).

The program requires that the data be in wide format. It must contain an earnings (or earnings residual) variable. If using a model with individual heterogeneity, it must also include an age (or labour market experience) variable. If using a model with cohort effects, it must also include a cohort indicator variable. The cohort indicator must be numeric and increase in increments of 1 [e.g A 4 cohort model could contain labels 1 to 4 or 1994 to 1997: either would work. However labels such as 1960, 1970, 1980 and 1990 would have to be recoded before being used.]

The program assumes that the name of the age (or labour market experience) variable is age; and that the time indicators on these variables run from 1 to T (T is `yearn'). If not, the user must enter these variables using the options.

If the time indicators do not increase by one in each successive period(eg if data are biannual), the user must manually create variables y1,y2,...,yT and age1,age2,...,ageT.

The program assumes that the name of the cohort variable is cohort; and that the cohort numbers run from 1 to C (C is `cohortn'). If not, the user must enter these names using the options.

The program uses the identity weighting matrix for GMM estimation (For discussion of the weighting matrix in GMM estimation see Altonji and Segal 1996) and the standard errors allow for unbalanced data using the approach reported in Haider(2001).

3.5 Saved Results

Scalars
e(numonent) number of moments used in estimating the model

Vectors and Matrices
e(b) coefficient vector
e(V) variance-covariance matrix of the estimators
e(moment_c) sample moments for earnings variable for cohorts 1 to Cohortn
e(perm_j) predicted permanent component of earnings variance for cohorts j=1..Cohortn

e(temp_j) predicted transitory component of earnings variance for cohorts j=1..Cohortn

4. Example

To illustrate the model with estimate the covariance structure of earnings using data taken from the 8 waves of the ECHP for Germany. The years covered by the survey are 1994-2001. The earnings variable are residuals from a first stage regression of earnings on age and age^2. This first stage must be carried out prior to analysis if required. The earnings variables were denoted as y1994 to y2001, and the individual age variables were denoted as potexp1994 to potexp2001. There were four cohorts labelled 1, 2, 3 and 4 and the cohort indicator variable was called cohort.

The following shows the results for a heterogeneous growth model with cohorts and an AR specification for the error term. In this example we see a negative relationship between α_i and β_i, indicating that people who start off with lower initial earnings grow faster. This is consistent with typical stories of investment in human capital models.

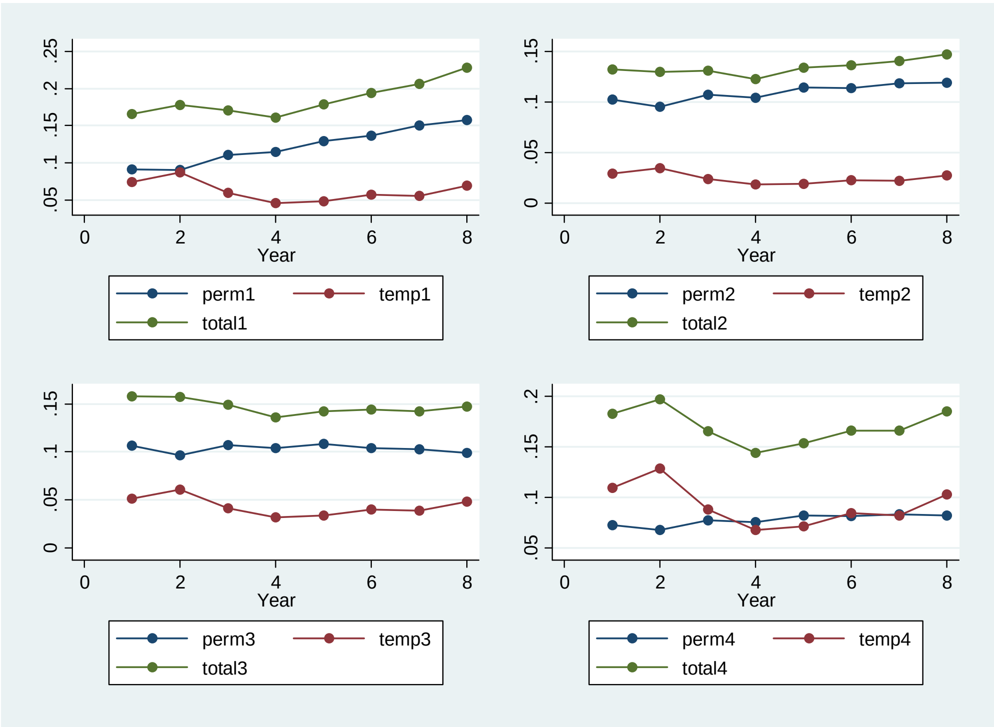
```
gmmcovearn y, yearn(8) modeln(7) cohortn(4) agevar(potexp) firstyr(1994)
(obs = 144)

Iteration 0:  residual SS =  .2195804
Iteration 1:  residual SS =  .1901487
Iteration 2:  residual SS =  .0911686
Iteration 3:  residual SS =  .0828604
Iteration 4:  residual SS =  .0696758
Iteration 5:  residual SS =  .0491634
Iteration 6:  residual SS =  .0071537
Iteration 7:  residual SS =  .0070074
Iteration 8:  residual SS =  .0070062
Iteration 9:  residual SS =  .0070061
Iteration 10:  residual SS =  .0070061
Iteration 11:  residual SS =  .0070061
Iteration 12:  residual SS =  .0070061

      Source |       SS       df       MS
-------------+------------------------------         Number of obs =       144
       Model |  2.16612446    26  .083312479         R-squared     =    0.9968
    Residual |  .007006143   118  .000059374         Adj R-squared =    0.9961
-------------+------------------------------         Root MSE      =  .0077055
       Total |  2.17313061   144  .015091185         Res. dev.     = -1021.378
```

moment	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
/sigalpha	.4386704	.0605218	7.25	0.000	.3188208	.55852
/rho	.3417545	.0361788	9.45	0.000	.2701106	.4133984
/sigv1	.0742455	.0062163	11.94	0.000	.0619355	.0865556
/sige	.0301694	.0112735	2.68	0.009	.0078448	.0524941
/l2	1.498369	.2427663	6.17	0.000	1.017626	1.979113
/l3	1.312377	.2569181	5.11	0.000	.8036097	1.821145
/l4	1.16253	.230131	5.05	0.000	.7068076	1.618252
/l5	1.193411	.2324543	5.13	0.000	.7330886	1.653734
/l6	1.297055	.2507745	5.17	0.000	.800453	1.793657
/l7	1.279364	.250722	5.10	0.000	.782866	1.775862
/l8	1.428737	.2790989	5.12	0.000	.8760449	1.981428
/p2	.9842997	.0222086	44.32	0.000	.9403207	1.028279
/p3	1.07534	.0254406	42.27	0.000	1.024961	1.125719
/p4	1.084179	.0279634	38.77	0.000	1.028804	1.139555
/p5	1.148706	.03183	36.09	0.000	1.085674	1.211738
/p6	1.168077	.033754	34.61	0.000	1.101235	1.234919
/p7	1.205651	.0364915	33.04	0.000	1.133388	1.277914
/p8	1.219382	.037295	32.70	0.000	1.145528	1.293236
/cp2	.989772	.0419597	23.59	0.000	.9066803	1.072864
/cp3	.7322496	.0536754	13.64	0.000	.6259578	.8385415
/cp4	.4866868	.0387748	12.55	0.000	.4099022	.5634714
/ct2	.6293472	.0447308	14.07	0.000	.540768	.7179265
/ct3	.8336289	.0377501	22.08	0.000	.7588735	.9083843
/ct4	1.214287	.0390211	31.12	0.000	1.137015	1.291559
/sigbeta	.0003872	.0000489	7.92	0.000	.0002904	.0004841
/covalphab~a	-.012158	.001729	-7.03	0.000	-.0155819	-.0087341

coefficients and corrected standard errors

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sigalpha	.4386704	.1706617	2.57	0.010	.1041796	.7731611
rho	.3417545	.0361378	9.46	0.000	.2709258	.4125833
sigv1	.0742455	.0118533	6.26	0.000	.0510135	.0974776
sige	.0301694	.0104324	2.89	0.004	.0097222	.0506166
l2	1.498369	.2208249	6.79	0.000	1.06556	1.931178
l3	1.312377	.2312365	5.68	0.000	.8591623	1.765593
l4	1.16253	.1944329	5.98	0.000	.7814482	1.543611
l5	1.193411	.1942604	6.14	0.000	.812668	1.574155
l6	1.297055	.2167253	5.98	0.000	.8722809	1.721829
l7	1.279364	.2119593	6.04	0.000	.8639312	1.694797
l8	1.428737	.2584138	5.53	0.000	.922255	1.935218
p2	.9842997	.0313718	31.38	0.000	.9228122	1.045787
p3	1.07534	.0433522	24.80	0.000	.9903709	1.160309
p4	1.084179	.0505351	21.45	0.000	.9851326	1.183226
p5	1.148706	.0694795	16.53	0.000	1.012529	1.284884
p6	1.168077	.078938	14.80	0.000	1.013362	1.322793
p7	1.205651	.0887035	13.59	0.000	1.031795	1.379507
p8	1.219382	.0922592	13.22	0.000	1.038557	1.400207
cp2	.989772	.1115841	8.87	0.000	.7710712	1.208473
cp3	.7322496	.1311764	5.58	0.000	.4751487	.9893506
cp4	.4866868	.1032264	4.71	0.000	.2843667	.6890069
ct2	.6293472	.0628975	10.01	0.000	.5060703	.7526241
ct3	.8336289	.0684399	12.18	0.000	.6994892	.9677685
ct4	1.214287	.091718	13.24	0.000	1.034523	1.394051
sigbeta	.0003872	.0001758	2.20	0.028	.0000426	.0007319
covalphabeta	-.012158	.0056141	-2.17	0.030	-.0231613	-.0011547

Below is a copy of the corresponding graph that would appear had the **graph(1)** option been specified.



5. Citations and conditions of Use.

This program is offered "as-is". Users should satisfy themselves that the program does what they want. Any bugs, comments or suggestions for improvement can b e-mailed to donal.oneill@nuim.ie.

Please Cite as:

Doris, A. D.O'Neill and O.Sweetman (2010) "GMMCOVEARN: A Stata Module for GMM Estimation of the Covariance Structure of Earnings," National University of Ireland, Maynooth.

Acknowledgements

We gratefully acknowledge the financial support provided by the Irish Research Council for the Humanities and Social Sciences

References

Altonji, J. and L. Segal (1996), 'Small-Sample Bias in GMM Estimation of Covariance Structures,' *Journal of Business & Economic Statistics,* Vol. 14, No. 3, pp. 353-366

Cappellari, L., (2004), 'The Dynamics and Inequality of Italian Men's Earnings: Long-Term Changes or Transitory Fluctuations?' *Journal of Human Resources*, Vol. 39(2), pp. 475-499.

Daly, M. and R. Valletta, (2008), 'Cross-National Trends in Earnings Inequality and Instability', *Economic Letters*, Vol. 99(2), pp. 215-219.

Dickens, R. (2000), 'The Evolution of Individual Male Earnings in Great Britain: 1975-95,' *Economic Journal*, Vol. 110, No. 460, pp. 27-49.

Doris, A. D.O'Neill and O.Sweetman (2010a), "Identification of the Covariance Structure of Earnings using the GMM Estimator," IZA Working paper no. 4952.

Doris, A, O'Neill, D and O.Sweetman (2010b), "Aggregate Earnings Inequality in Europe: Permanent Differences or Transitory Fluctuations?" NUIM Working Paper N211-10

Haider, S., (2001), 'Earnings Instability and Earnings Inequality of Males in the United States: 1967-1991', *Journal of Labor Economics*, Vol. 19(4), pp. 799-836.

Moffitt, R. and P. Gottschalk (1995), 'Trends in the Autocovariance Structure of Earnings in the U.S., 1969-1987,' mimeo, Johns Hopkins University.

Moffitt, R. and P. Gottschalk, (2002), 'Trends in the Transitory Variance of Earnings in the United States,' *Economic Journal*, Vol. 112, pp. 68-73.

Moffitt, R. and P. Gottschalk, (2008), 'Trends in the Transitory Variance of Male Earnings in the U.S., 1970-2004', Boston College, Working Paper.

Ramos, X., (2003), 'The Covariance Structure of Earnings in Great Britain, 1991-1999', *Economica*, Vol. 70, pp. 353-374.