

# Longitudinal Data Analysis

## Instructor: Natasha Sarkisian

### Panel Data Analysis: Random Effects Models, GEE Models, and GLS Models

#### Random Effects Models

In fixed effects models, each dummy variable removes one degree of freedom from the model; thus, fixed effects models work well when you have a substantial number of time periods. To avoid losing the degrees of freedom, we can estimate random effects models. The model still decomposes the residuals:  $Y_{it} = \alpha + X_{it}\beta + u_i + e_{it}$  where  $u_i$  represents the effect of unit  $i$  and  $e_{it}$  is the residual effect for time point  $t$  for that unit. But in a random effects model, unit residuals  $u_i$  do not have specific values –  $u_i$  is a normally distributed random variable (hence the name – random effects).

The nature of the coefficients  $\beta$  also changes as we go from a fixed effects to a random effects model – in a random effects model, we are not only predicting change over time but also explaining the differences among the units. Thus, the data on cross-sectional variation are utilized in estimating independent variables' effects. Because the predictors are used to explain not only change over time but also differences among units, the random unit residual variable  $u$  is assumed to be uncorrelated with  $X\beta$ :  $\text{corr}(u_i, X\beta) = 0$

```
. xtreg v41 v76 v5 v7 v19, re
```

```
Random-effects GLS regression           Number of obs   =       4945
Group variable (i): v1                 Number of groups =        124

R-sq:  within = 0.1782                 Obs per group:  min =         1
      between = 0.0000                   avg   =       39.9
      overall  = 0.0820                   max   =       148

Random effects u_i ~ Gaussian          Wald chi2(4)     =       979.05
corr(u_i, X) = 0 (assumed)             Prob > chi2      =       0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
v41						
v76	.0399844	.0115547	3.46	0.001	.0173376	.0626311
v5	-.0356356	.0060776	-5.86	0.000	-.0475474	-.0237238
v7	.5018372	.0721678	6.95	0.000	.3603908	.6432835
v19	.0074809	.0003227	23.19	0.000	.0068485	.0081133
_cons	-6.525467	21.13472	-0.31	0.758	-47.94875	34.89781
sigma_u	190.70328					
sigma_e	242.31081					
rho	.38248708	(fraction of variance due to u_i)				

Note that the variance decomposition is similar to that in the fixed effects model, but a significance test for unit-level variance is not included. But we can easily obtain it:

```
. xttest0
Breusch and Pagan Lagrangian multiplier test for random effects:
v41[v1,t] = Xb + u[v1] + e[v1,t]
```

```
Estimated results:
-----+-----
          |          Var          sd = sqrt(Var)
-----+-----
    v41 |    151964.3         389.826
         |    58714.53         242.3108
         |    36367.74         190.7033
```

```
Test:   Var(u) = 0
          chi2(1) = 93324.92
          Prob > chi2 = 0.0000
```

Thus, we reject the null hypothesis that country-specific residuals are all zero – there is a significant amount of variance among countries above and beyond that explained by our predictors.

So far we estimated our model using GLS (generalized least squares) estimation method; we could also estimate the same model using maximum likelihood estimation option:

```
. xtreg v41 v76 v5 v7 v19, re mle
Fitting constant-only model:
Iteration 0:   log likelihood = -34876.131
Iteration 1:   log likelihood = -34857.061
Iteration 2:   log likelihood = -34836.518
Iteration 3:   log likelihood = -34831.726
Iteration 4:   log likelihood = -34831.221
Iteration 5:   log likelihood = -34831.212
Iteration 6:   log likelihood = -34831.212
Fitting full model:
Iteration 0:   log likelihood = -34417.656
Iteration 1:   log likelihood = -34385.528
Iteration 2:   log likelihood = -34377.18
Iteration 3:   log likelihood = -34376.042
Iteration 4:   log likelihood = -34376
Iteration 5:   log likelihood = -34376
Random-effects ML regression
Group variable: v1
Number of obs   = 4945
Number of groups = 124

Random effects u_i ~ Gaussian
Obs per group:  min = 1
                avg = 39.9
                max = 148

LR chi2(4)      = 910.42
Prob > chi2     = 0.0000

Log likelihood = -34376
LR chi2(4)      = 910.42
Prob > chi2     = 0.0000
-----+-----
          |          Coef.          Std. Err.          z          P>|z|          [95% Conf. Interval]
-----+-----
    v76 |    .0415761    .0116619         3.57    0.000         .0187192         .0644329
     v5 |   -.0446729    .0067226        -6.65    0.000        -.0578491        -.0314968
     v7 |    .5321515    .0734379         7.25    0.000         .3882158         .6760871
    v19 |    .0076527    .0003252        23.53    0.000         .0070153         .0082901
   _cons |  -3.086056    25.69899         -0.12    0.904        -53.45514         47.28303
-----+-----
  /sigma_u |    250.7508    17.72269         218.3136    288.0076
  /sigma_e |    242.4077     2.471851         237.6111    247.3012
     rho   |    .5169129     .0357856         .4468744     .586432
-----+-----
Likelihood-ratio test of sigma_u=0:  chibar2(01)= 3742.21  Prob>=chibar2 = 0.000
```

The same model can be fit using `xtmixed` command – we will later use this command for mixed model, and the random effects model is a basic case of such a model:

```
. xtmixed v41 v76 v5 v7 v19 || v1:
```

```
Performing EM optimization:
```

```
Performing gradient-based optimization:
```

```
Iteration 0: log restricted-likelihood = -34388.795
```

```
Iteration 1: log restricted-likelihood = -34388.795
```

```
Computing standard errors:
```

```
Mixed-effects REML regression          Number of obs      =      4945
Group variable: v1                     Number of groups   =      124

Obs per group: min =          1
                  avg =         39.9
                  max =         148
```

```
Log restricted-likelihood = -34388.795      Wald chi2(4)      =    1015.07
                                           Prob > chi2      =      0.0000
```

v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
v76	.0416042	.0116643	3.57	0.000	.0187426 .0644658
v5	-.0448407	.0063998	-7.01	0.000	-.057384 -.0322974
v7	.5326954	.0731793	7.28	0.000	.3892666 .6761243
v19	.0076559	.0003229	23.71	0.000	.007023 .0082888
_cons	-3.018167	25.81173	-0.12	0.907	-53.60823 47.5719

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
v1: Identity			
sd(_cons)	252.2981	17.89278	219.5571 289.9216
sd(Residual)	242.4974	2.473587	237.6974 247.3943

```
LR test vs. linear regression: chibar2(01) = 3746.44 Prob >= chibar2 = 0.0000
```

As mentioned above, random effects coefficients have a dual nature: They simultaneously explain change over time and the cross-sectional differences among units. The implicit assumption is that both types of effects are the same. That is, when we say that a one unit increase in X is associated with a b units increase in Y, a one unit increase might mean two things:

1. We observe two different countries with a one unit difference in X between them.
2. We observe one country, and its X value increases by one unit.

In a random effects model, we are assuming that both of those produce the same effect on Y. That is, for instance, we assume that if two countries have different population density by 1 unit, and if for a given country the population density increases by one unit, the effect of military size would be the same.

We test this assumption using the Hausman test. The Hausman test checks a more efficient model against a less efficient but consistent model to make sure that the more efficient model also gives consistent results. The null hypothesis is that the coefficients estimated by the efficient random effects estimator are the same as the ones estimated by the consistent fixed effects estimator. If they are, then it is safe to use a random effects model. If the two sets of coefficients are significantly different, then the random effects model is problematic. It is best to use hausman test with sigmamore option; it avoids problems with the matrix  $[V_b - V_B]$  not being positive definite.

```
. qui xtreg v41 v76 v5 v7 v19, fe
. est store fixed
. qui xtreg v41 v76 v5 v7 v19, re
. est store random
. hausman fixed random, sigmamore
```

	---- Coefficients ----		(b-B)	sqrt(diag(V_b-V_B))
	(b)	(B)	Difference	S.E.
	fixed	random		
v76	.0445623	.0399844	.0045779	.0035381
v5	-.0637468	-.0356356	-.0281112	.0037499
v7	.5901123	.5018372	.0882752	.02503
v19	.0080274	.0074809	.0005465	.0000754

b = consistent under Ho and Ha; obtained from xtreg  
 B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(4) = (b-B)'[(V\_b-V\_B)^(-1)](b-B)  
 = 81.09  
 Prob>chi2 = 0.0000

In this case, we reject the null hypothesis – fixed effects and random effects coefficients are significantly different. Indeed, this assumption more often holds for large longitudinal survey data sets, and less often holds for smaller cross-sectional samples, for example, a sample of countries. Examining the coefficients, we might suspect that v7 or v5 are responsible.

To better understand the meaning of the Hausman test, let's introduce the between effects model.

```
. xtreg v41 v76 v5 v7 v19, be
```

```
Between regression (regression on group means) Number of obs = 4945
Group variable (i): v1 Number of groups = 124

R-sq: within = 0.0295 Obs per group: min = 1
      between = 0.1300 avg = 39.9
      overall = 0.0232 max = 148

sd(u_i + avg(e_i.))= 202.0667 F(4,119) = 4.44
Prob > F = 0.0022
```

v41	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
v76	.0085641	.0399603	0.21	0.831	-.0705613 .0876895
v5	.03787	.0118493	3.20	0.002	.0144073 .0613327
v7	.2871046	.2349388	1.22	0.224	-.1780978 .7523069
v19	-.0003675	.0014284	-0.26	0.797	-.0031959 .0024608
_cons	13.05287	41.68082	0.31	0.755	-69.47932 95.58507

This type of analysis is equivalent to taking the mean of each variable across time for each case and running a regression on the collapsed dataset of means. As this results in a loss of information, between effects are rarely used. The between effects estimator is mostly important because Stata's random-effects estimator is a weighted average of a fixed effects and a between effects coefficient. Thus, implicitly, the Hausman test assesses whether fixed effects and between effects produce the same coefficients. If they do, it is appropriate to combine them into a random effects model. Comparing these coefficients to the fixed effects coefficients in the Hausman output, we see some major differences for v5 and v7. We could also estimate the two types of effects (over time and across units) separately in a single random effects model:

```
. for var v76 v5 v7 v19: egen Xmean=mean(X), by(v1)

-> egen v76mean=mean(v76), by(v1)
(775 missing values generated)

-> egen v5mean=mean(v5), by(v1)
(50 missing values generated)

-> egen v7mean=mean(v7), by(v1)
(665 missing values generated)

-> egen v19mean=mean(v19), by(v1)
(448 missing values generated)

. for var v76 v5 v7 v19: gen Xchange=X-Xmean

-> gen v76change=v76-v76mean
(2956 missing values generated)

-> gen v5change=v5-v5mean
(161 missing values generated)

-> gen v7change=v7-v7mean
(1166 missing values generated)

-> gen v19change=v19-v19mean
(2643 missing values generated)

. xtreg v41 v76mean v76change v5mean v5change v7mean v7change v19mean v19change, re

Random-effects GLS regression                Number of obs      =       4945
Group variable (i): v1                      Number of groups   =        124

R-sq:   within = 0.1806                    Obs per group:    min =         1
         between = 0.1550                    avg =        39.9
         overall = 0.1101                    max =        148

Random effects u_i ~ Gaussian                Wald chi2(8)       =    1081.28
corr(u_i, X) = 0 (assumed)                  Prob > chi2        =         0.0000
```

v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
v76mean	-.002741	.0410229	-0.07	0.947	-.0831443 .0776624	
v76change	.045371	.0119146	3.81	0.000	.0220189 .0687231	
v5mean	.0487256	.0127597	3.82	0.000	.023717 .0737342	
v5change	-.0637998	.0070639	-9.03	0.000	-.0776448 -.0499549	
v7mean	.1582229	.2472661	0.64	0.522	-.3264097 .6428556	
v7change	.591951	.0752684	7.86	0.000	.4444275 .7394744	
v19mean	-.000504	.0016745	-0.30	0.763	-.003786 .002778	
v19change	.0080235	.0003285	24.43	0.000	.0073797 .0086672	
_cons	31.74789	41.07776	0.77	0.440	-48.76303 112.2588	
sigma_u	188.7826					
sigma_e	242.33597					
rho	.37766805	(fraction of variance due to u_i)				

```

. test v76mean=v76change

( 1)  v76mean - v76change = 0

      chi2( 1) =    1.27
      Prob > chi2 =    0.2591

. test v5mean=v5change

( 1)  v5mean - v5change = 0

      chi2( 1) =   56.86
      Prob > chi2 =    0.0000

. test v7mean=v7change

( 1)  v7mean - v7change = 0

      chi2( 1) =    2.81
      Prob > chi2 =    0.0937

. test v19mean=v19change

( 1)  v19mean - v19change = 0

      chi2( 1) =   24.94
      Prob > chi2 =    0.0000

```

Thus, there are really two kinds of information in cross-sectional time-series data:

1. The cross-sectional information reflected in the differences among units.
2. The time-series or within-unit information reflected in the changes within units.

A between effects model uses only the cross-sectional information and asks: “What is the expected difference in Y between two countries that differ by 1 in X?”, while a fixed effects model uses only the time-series information and asks, “What is the expected change in a country’s value of Y if its value of X increases by 1?” A random effects model combines those two questions, but really, it may turn out that the answers to those two questions are the same or they may be different. If they are different, we could either use a fixed effects model, or we can separate the two types of effects within a random effects model, but we should be able to explain why the effects are different. Statistically, a fixed effects model is always a reasonable thing to



```
H0: no first-order autocorrelation
      F( 1, 119) = 31.401
      Prob > F = 0.0000
```

Here, the hypothesis of no first order autocorrelation is rejected; therefore, we would want a model explicitly accounting for autoregressive error term. We can use xtregar models that assume that:

$$y_{it} = a + x_{it} * B + u_i + e_{it}$$

where  $e_{it} = \rho * e_{i,t-1} + z_{it}$  with  $|\rho| < 1$

```
. xtregar v41 v76 v5 v7 v19, fe

FE (within) regression with AR(1) disturbances   Number of obs   =   4821
Group variable (i): v1                          Number of groups =   122

R-sq:  within = 0.0088                          Obs per group:  min =    1
      between = 0.0054                             avg   =   39.5
      overall = 0.0381                             max   =   147

corr(u_i, Xb) = -0.1084                          F(4,4695)       =   10.42
                                                    Prob > F        =   0.0000
```

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
v41						
v76	.0375775	.0163388	2.30	0.021	.0055459	.0696091
v5	-.0306358	.0257023	-1.19	0.233	-.0810244	.0197529
v7	.2811527	.1152901	2.44	0.015	.05513	.5071755
v19	.0019622	.0003692	5.31	0.000	.0012383	.0026861
_cons	70.06587	1.590655	44.05	0.000	66.94744	73.1843
rho_ar	.97935195					
sigma_u	231.59613					
sigma_e	56.802909					
rho_fov	.94325746	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(121,4695) = 0.97      Prob > F = 0.5709
```

```
. xtregar v41 v76 v5 v7 v19, re

RE GLS regression with AR(1) disturbances   Number of obs   =   4945
Group variable (i): v1                          Number of groups =   124

R-sq:  within = 0.1460                          Obs per group:  min =    1
      between = 0.0552                             avg   =   39.9
      overall = 0.0852                             max   =   148

corr(u_i, Xb) = 0 (assumed)                      Wald chi2(5)    =   40.96
                                                    Prob > chi2    =   0.0000
```

----- theta -----				
min	5%	median	95%	max
0.0563	0.0627	0.1017	0.1118	0.1313

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
v41						
v76	.023926	.0138356	1.73	0.084	-.0031912	.0510433

```

      v5 | .0052999 .0110826 0.48 0.632 -.0164216 .0270214
      v7 | .1931183 .0985973 1.96 0.050 -.0001289 .3863654
      v19 | .0016473 .0003285 5.01 0.000 .0010034 .0022913
      _cons | 44.1919 30.49233 1.45 0.147 -15.57198 103.9558
-----+-----
      rho_ar | .97935195 (estimated autocorrelation coefficient)
      sigma_u | 97.247728
      sigma_e | 56.10666
      rho_fov | .75026276 (fraction of variance due to u_i)
-----+-----

```

Xtregar also offers additional tests for autocorrelation, based on Durbin-Watson statistic.

```
. xtregar v41 v76 v5 v7 v19, re lbi
```

```

RE GLS regression with AR(1) disturbances      Number of obs      =      4945
Group variable: v1                            Number of groups    =      124

R-sq:  within = 0.1460                        Obs per group:  min =      1
      between = 0.0552                          avg =     39.9
      overall  = 0.0852                          max =     148

corr(u_i, Xb)      = 0 (assumed)                Wald chi2(5)        =     40.96
                                                           Prob > chi2         =     0.0000

```

```

-----+----- theta -----+-----
      min      5%      median      95%      max
0.0563  0.0627  0.1017  0.1118  0.1313

```

```

-----+-----
      v41 |      Coef.  Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
      v76 |      .023926  .0138356     1.73  0.084     -0.0031912     .0510433
      v5  |      .0052999  .0110826     0.48  0.632     -0.0164216     .0270214
      v7  |      .1931183  .0985973     1.96  0.050     -0.0001289     .3863654
      v19 |      .0016473  .0003285     5.01  0.000     .0010034     .0022913
      _cons |      44.1919  30.49233     1.45  0.147     -15.57198     103.9558
-----+-----
      rho_ar | .97935195 (estimated autocorrelation coefficient)
      sigma_u | 97.247728
      sigma_e | 56.10666
      rho_fov | .75026276 (fraction of variance due to u_i)
-----+-----

```

```

modified Bhargava et al. Durbin-Watson = .08749333
Baltagi-Wu LBI = .13942445

```

A value of the modified Durbin-Watson statistic or Baltagi-Wu LBI-statistic of 2 indicates no autocorrelation (the values lie between 0 and 4). As a rough rule of thumb, values below 1 mean you should definitely correct for serial correlation. Small values indicate successive error terms are positively correlated. With statistic values  $>2$ , successive error terms are, on average, much different in value to one another, i.e., negatively correlated. In regressions, this can imply an underestimation of the level of statistical significance.

## GEE models

So far we have dealt with so-called unit-specific models where the residuals were separated into two components, unit residual and time-specific residual. But another group of models, while taking into account that groups will have similar residuals, does not separate them but rather explicitly models the structure of error term covariances. Population-averaged xtreg model is an example of such a model:

```
. xtreg v41 v76 v5 v7 v19, pa
Iteration 1: tolerance = .94544759
Iteration 2: tolerance = .0706797
Iteration 3: tolerance = .00002357
Iteration 4: tolerance = 7.308e-09
GEE population-averaged model
Group variable:                v1
Link:                          identity
Family:                        Gaussian
Correlation:                   exchangeable
Scale parameter:               148090.7
Number of obs                  =      4945
Number of groups               =      124
Obs per group: min            =         1
                             avg            =      39.9
                             max            =      148
Wald chi2(4)                  =    1681.71
Prob > chi2                    =      0.0000
```

	v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
v76		.0433727	.0093337	4.65	0.000	.0250789	.0616665
v5		-.0558803	.0053575	-10.43	0.000	-.0663807	-.0453798
v7		.5671214	.0588283	9.64	0.000	.4518201	.6824227
v19		.0078699	.0002567	30.66	0.000	.0073668	.0083729
_cons		1.67636	31.6295	0.05	0.958	-60.31632	63.66904

Note that this model makes the same assumption about the relationship between between effects and fixed effects. Xtreg, pa is a special case of GEE (generalized estimating equations) model – let’s compare:

```
. xtgee v41 v76 v5 v7 v19
Iteration 1: tolerance = .94544759
Iteration 2: tolerance = .0706797
Iteration 3: tolerance = .00002357
Iteration 4: tolerance = 7.308e-09
GEE population-averaged model
Group variable:                v1
Link:                          identity
Family:                        Gaussian
Correlation:                   exchangeable
Scale parameter:               148090.7
Number of obs                  =      4945
Number of groups               =      124
Obs per group: min            =         1
                             avg            =      39.9
                             max            =      148
Wald chi2(4)                  =    1681.71
Prob > chi2                    =      0.0000
```

	v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
v76		.0433727	.0093337	4.65	0.000	.0250789	.0616665
v5		-.0558803	.0053575	-10.43	0.000	-.0663807	-.0453798
v7		.5671214	.0588283	9.64	0.000	.4518201	.6824227
v19		.0078699	.0002567	30.66	0.000	.0073668	.0083729
_cons		1.67636	31.6295	0.05	0.958	-60.31632	63.66904

This model, however, can incorporate a wide range of covariance structures:

Independent = OLS regression:

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$
$t_1$	—	0	0	0	0	0
$t_2$	0	—	0	0	0	0
$t_3$	0	0	—	0	0	0
$t_4$	0	0	0	—	0	0
$t_5$	0	0	0	0	—	0
$t_6$	0	0	0	0	0	—

Autoregressive:

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$
$t_1$	—	$\rho^1$	$\rho^2$	$\rho^3$	$\rho^4$	$\rho^5$
$t_2$	$\rho^1$	—	$\rho^1$	$\rho^2$	$\rho^3$	$\rho^4$
$t_3$	$\rho^2$	$\rho^1$	—	$\rho^1$	$\rho^2$	$\rho^3$
$t_4$	$\rho^3$	$\rho^2$	$\rho^1$	—	$\rho^1$	$\rho^2$
$t_5$	$\rho^4$	$\rho^3$	$\rho^2$	$\rho^1$	—	$\rho^1$
$t_6$	$\rho^5$	$\rho^4$	$\rho^3$	$\rho^2$	$\rho^1$	—

Exchangeable = xtreg with pa option:

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$
$t_1$	—	$\rho$	$\rho$	$\rho$	$\rho$	$\rho$
$t_2$	$\rho$	—	$\rho$	$\rho$	$\rho$	$\rho$
$t_3$	$\rho$	$\rho$	—	$\rho$	$\rho$	$\rho$
$t_4$	$\rho$	$\rho$	$\rho$	—	$\rho$	$\rho$
$t_5$	$\rho$	$\rho$	$\rho$	$\rho$	—	$\rho$
$t_6$	$\rho$	$\rho$	$\rho$	$\rho$	$\rho$	—

Nonstationary:

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$
$t_1$	—	$\rho_1$	$\rho_2$	0	0	0
$t_2$	$\rho_1$	—	$\rho_3$	$\rho_4$	0	0
$t_3$	$\rho_2$	$\rho_3$	—	$\rho_5$	$\rho_6$	0
$t_4$	0	$\rho_4$	$\rho_5$	—	$\rho_7$	$\rho_8$
$t_5$	0	0	$\rho_6$	$\rho_7$	—	$\rho_9$
$t_6$	0	0	0	$\rho_8$	$\rho_9$	—

Stationary:

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$
$t_1$	—	$\rho_1$	$\rho_2$	0	0	0
$t_2$	$\rho_1$	—	$\rho_1$	$\rho_2$	0	0
$t_3$	$\rho_2$	$\rho_1$	—	$\rho_1$	$\rho_2$	0
$t_4$	0	$\rho_2$	$\rho_1$	—	$\rho_1$	$\rho_2$
$t_5$	0	0	$\rho_2$	$\rho_1$	—	$\rho_1$
$t_6$	0	0	0	$\rho_2$	$\rho_1$	—

Unstructured:

	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$
$t_1$	—	$\rho_1$	$\rho_2$	$\rho_3$	$\rho_4$	$\rho_5$
$t_2$	$\rho_1$	—	$\rho_6$	$\rho_7$	$\rho_8$	$\rho_9$
$t_3$	$\rho_2$	$\rho_6$	—	$\rho_{10}$	$\rho_{11}$	$\rho_{12}$
$t_4$	$\rho_3$	$\rho_7$	$\rho_{10}$	—	$\rho_{13}$	$\rho_{14}$
$t_5$	$\rho_4$	$\rho_8$	$\rho_{11}$	$\rho_{13}$	—	$\rho_{15}$
$t_6$	$\rho_5$	$\rho_9$	$\rho_{12}$	$\rho_{14}$	$\rho_{15}$	—

To model them in GEE, we use correlation option with the following specifications:

exchangeable	exchangeable
independent	independent
unstructured	unstructured
fixed matname	user-specified
ar #	autoregressive of order #
stationary #	stationary of order #
nonstationary #	nonstationary of order #

Note that for autoregressive, stationary, and nonstationary structures, you can specify order – i.e., the distance after which correlations should be assumed zero.

Allowed characteristics of the data for each correlation structure:

```

--characteristics allowed--
                Unequal
Correlation      Unbalanced spacing Gaps
-----
independent      yes          yes     yes
exchangeable     yes          yes     yes
ar k             yes (*)     no      no
stationary k     yes (*)     no      no
nonstationary k yes (*)     no      no
unstructured     yes          yes     yes
fixed            yes          yes     yes
-----
(*) All panels must have at least k+1 obs.

```

Definitions:

1. Panels are balanced if each has the same number of observations.
2. Panels are equally spaced if the interval between observations is constant.
3. Panels have gaps if some observations are missing.

Here, we have unequal spacing and gaps problems, so we can only estimate independent or exchangeable structures (in theory, we could estimate unstructured as well, but it does not converge for our example).

As mentioned above, OLS model falls under GEE models as a special case (with independent correlation structure):

```
. xtgee v41 v76 v5 v7 v19, corr(indep)
```

Iteration 1: tolerance = 8.385e-15

```

GEE population-averaged model      Number of obs      =      4945
Group variable:                    v1                 Number of groups   =       124
Link:                               identity            Obs per group: min =         1
Family:                            Gaussian              avg =       39.9
Correlation:                       independent         max =       148
                                     Wald chi2(4)       =     569.33
Scale parameter:                   136247.2           Prob > chi2        =     0.0000

Pearson chi2(4945):                6.737e+08          Deviance           = 6.737e+08
Dispersion (Pearson):              136247.2           Dispersion         = 136247.2

```

```

-----
      v41 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      v76 |   .0107072   .0107906     0.99   0.321    - .010442   .0318564
       v5 |   .0038175   .0032191     1.19   0.236    - .0024918   .0101268
       v7 |   .2710273   .0596875     4.54   0.000     .1540421   .3880126
      v19 |   .0073414   .0004121    17.81   0.000     .0065337   .0081491
     _cons |  44.82276   11.09903     4.04   0.000     23.06906   66.57646
-----

```

We can see the matrix of correlations using “estat mcor” after xtgee (here, it’s a matrix of zeros).

How do we decide what to use? You can estimate the OLS model and examine correlations among residuals (after generating residuals, you should reshape them into wide and then examine correlations among residuals from different time points and see if you can identify any pattern). You can also use the autocorrelation tests described above as well as heterogeneity test that we will discuss soon.

## GLS Models

Error term covariance structures can be modeled not only in population average models but also in GLS models. In addition to modeling correlations among time points within individuals, GLS models also allow us to allow for error term heterogeneity across panels. They use feasible generalized least squares and allow for AR(1) autocorrelation within panels and cross-sectional correlation and heteroskedasticity across panels. Once again, OLS model can be considered as a special case of a GLS model with homoskedastic panels and no autocorrelation:

```
. xtglm v41 v76 v5 v7 v19
```

```
Cross-sectional time-series FGLS regression
```

```
Coefficients:  generalized least squares
Panels:       homoskedastic
Correlation:  no autocorrelation
```

```
Estimated covariances      =          1      Number of obs      =       4945
Estimated autocorrelations =          0      Number of groups   =        124
Estimated coefficients      =          5      Obs per group: min =          1
                                           avg =   39.87903
                                           max =         148
                                           Wald chi2(4)      =       569.33
                                           Prob > chi2       =       0.0000
```

v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
v76	.0107072	.0107906	0.99	0.321	-.010442 .0318564
v5	.0038175	.0032191	1.19	0.236	-.0024918 .0101268
v7	.2710273	.0596875	4.54	0.000	.1540421 .3880126
v19	.0073414	.0004121	17.81	0.000	.0065337 .0081491
_cons	44.82276	11.09903	4.04	0.000	23.06906 66.57646

### Options:

```
panels(iid)           use i.i.d. error structure
panels(heteroskedastic) use heteroskedastic but uncorrelated error structure
panels(correlated)    use heteroskedastic and correlated error structure
corr(independent)     use independent autocorrelation structure
corr(ar1)             use AR1 autocorrelation structure
corr(psar1)          use panel-specific AR1 autocorrelation structure
```

```
. xtglm v41 v76 v5 v7 v19, corr(ar1)
```

```
(note: 2 observations dropped because only 1 obs in group)
v2 is not regularly spaced or does not have intervals of delta -- use
the force option to treat the intervals as though they were regular
r(459);
```

Same problem with spacing, but we will force it:

```
. xtglm v41 v76 v5 v7 v19, corr(ar1) force
```

```
(note: 2 observations dropped because only 1 obs in group)
```

```
Cross-sectional time-series FGLS regression
```

```
Coefficients:  generalized least squares
Panels:       homoskedastic
Correlation:  common AR(1) coefficient for all panels (1.0110)
```

```

Estimated covariances      =          1      Number of obs      =      4943
Estimated autocorrelations =          1      Number of groups   =      122
Estimated coefficients      =          5      Obs per group: min =         2
                                           avg =    40.51639
                                           max =         148
                                           Wald chi2(4)      =    124.63
                                           Prob > chi2       =     0.0000

```

```

-----+-----
      v41 |      Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      v76 |  -.0061443   .0149567   -0.41  0.681   - .0354589   .0231703
       v5 |  -.1039554   .0294458   -3.53  0.000   - .161668   -.0462427
       v7 |   .3623291   .1120681    3.23  0.001    .1426796   .5819786
      v19 |   .0047278   .0004632   10.21  0.000    .0038199   .0056358
     _cons |  158.3503    99.37225    1.59  0.111   -36.4157    353.1164
-----+-----

```

Let's consider the issue of heteroskedastic error structures:

```
. xtgls v41 v76 v5 v7 v19, panels(hetero)
```

Cross-sectional time-series FGLS regression

```

Coefficients:  generalized least squares
Panels:        heteroskedastic
Correlation:   no autocorrelation

```

```

Estimated covariances      =          124      Number of obs      =      4945
Estimated autocorrelations =          0      Number of groups   =      124
Estimated coefficients      =          5      Obs per group: min =         1
                                           avg =    39.87903
                                           max =         148
                                           Wald chi2(4)      =   1144.58
                                           Prob > chi2       =     0.0000

```

```

-----+-----
      v41 |      Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      v76 |  -.0032924   .0027694   -1.19  0.234   - .0087203   .0021354
       v5 |   .0170866   .0012755   13.40  0.000    .0145866   .0195866
       v7 |   .2857526   .0191251   14.94  0.000    .248268    .3232372
      v19 |   .0024015   .0002565    9.36  0.000    .0018988   .0029042
     _cons |  15.58696    2.087263    7.47  0.000    11.496    19.67792
-----+-----

```

To test whether the model with heteroskedastic error structures is superior, we will reestimate the same model using iterative GLS and do a log-likelihood test:

```

. xtgls v41 v76 v5 v7 v19, pan(hetero) igls
Iteration 1: tolerance = .53625946
...
Iteration 51: tolerance = 9.492e-08

```

Cross-sectional time-series FGLS regression

```

Coefficients:  generalized least squares
Panels:        heteroskedastic
Correlation:   no autocorrelation

```

```

Estimated covariances      =          124      Number of obs      =      4945
Estimated autocorrelations =          0      Number of groups   =      124

```

```

Estimated coefficients      =          5          Obs per group: min =          1
                                                                    avg = 39.87903
                                                                    max = 148
Log likelihood              = -24208.91      Wald chi2(4) = 17494.30
                                                                    Prob > chi2  = 0.0000

```

v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
v76	-7.17e-06	4.27e-07	-16.80	0.000	-8.01e-06 -6.34e-06
v5	.0005365	9.85e-06	54.49	0.000	.0005172 .0005558
v7	.0336151	.0002944	114.18	0.000	.033038 .0341921
v19	3.59e-07	8.11e-08	4.42	0.000	2.00e-07 5.18e-07
_cons	.9862699	.0003808	2589.95	0.000	.9855236 .9870163

```
. est store hetero
```

```
. xtgls v41 v76 v5 v7 v19, igls
Iteration 1: tolerance = 0
```

Cross-sectional time-series FGLS regression

```

Coefficients: generalized least squares
Panels:       homoskedastic
Correlation:  no autocorrelation

```

```

Estimated covariances      =          1          Number of obs      =          4945
Estimated autocorrelations =          0          Number of groups   =          124
Estimated coefficients      =          5          Obs per group: min =          1
                                                                    avg = 39.87903
                                                                    max = 148
                                                                    Wald chi2(4) = 569.33
Log likelihood              = -36247.11      Prob > chi2 = 0.0000

```

v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
v76	.0107072	.0107906	0.99	0.321	-.010442 .0318564
v5	.0038175	.0032191	1.19	0.236	-.0024918 .0101268
v7	.2710273	.0596875	4.54	0.000	.1540421 .3880126
v19	.0073414	.0004121	17.81	0.000	.0065337 .0081491
_cons	44.82276	11.09903	4.04	0.000	23.06906 66.57646

```
. local df = e(N_g) - 1
```

```
. lrtest hetero , df(`df')
```

```

Likelihood-ratio test          LR chi2(123)= 24076.39
(Assumption: . nested in hetero) Prob > chi2 = 0.0000

```

Heterogenous error structure is far superior. But, as we can see from its relationship to OLS, by default xtgls does not include either random or fixed effects to control for variation in the level of our dependent variable across units. If we want, we could introduce dummies to model fixed effects:

```

. xi: xtgls v41 v76 v5 v7 v19 i.v1, pan(hetero)
i.v1          _Iv1_10-1300      (naturally coded; _Iv1_10 omitted)
note: _Iv1_51 dropped because of collinearity
note: _Iv1_65 dropped because of collinearity
note: _Iv1_69 dropped because of collinearity

```

```

note: _Iv1_70 dropped because of collinearity
note: _Iv1_230 dropped because of collinearity
note: _Iv1_360 dropped because of collinearity
note: _Iv1_421 dropped because of collinearity
note: _Iv1_422 dropped because of collinearity
note: _Iv1_423 dropped because of collinearity
note: _Iv1_424 dropped because of collinearity
note: _Iv1_425 dropped because of collinearity
note: _Iv1_426 dropped because of collinearity
note: _Iv1_428 dropped because of collinearity
note: _Iv1_429 dropped because of collinearity
note: _Iv1_571 dropped because of collinearity
note: _Iv1_572 dropped because of collinearity
note: _Iv1_573 dropped because of collinearity
note: _Iv1_574 dropped because of collinearity
note: _Iv1_575 dropped because of collinearity
note: _Iv1_576 dropped because of collinearity
note: _Iv1_630 dropped because of collinearity
note: _Iv1_631 dropped because of collinearity
note: _Iv1_760 dropped because of collinearity
note: _Iv1_780 dropped because of collinearity
note: _Iv1_901 dropped because of collinearity
note: _Iv1_965 dropped because of collinearity
note: _Iv1_1020 dropped because of collinearity
note: _Iv1_1121 dropped because of collinearity
note: _Iv1_1185 dropped because of collinearity
note: _Iv1_1215 dropped because of collinearity
note: _Iv1_1261 dropped because of collinearity
note: _Iv1_1291 dropped because of collinearity

```

Cross-sectional time-series FGLS regression

```

Coefficients: generalized least squares
Panels:      heteroskedastic
Correlation: no autocorrelation

```

```

Estimated covariances      =      124      Number of obs      =      4945
Estimated autocorrelations =      0      Number of groups   =      124
Estimated coefficients     =      138      Obs per group: min =      1
                                           avg = 39.87903
                                           max = 148
                                           Wald chi2(125)    = 39749.41
                                           Prob > chi2      = 0.0000

```

```

-----
      v41 |      Coef.  Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
      v76 |   .0142667   .0022677     6.29   0.000   .0098221   .0187112
      v5  |  -.0044016   .0018232    -2.41   0.016  -.007975  -.0008281
      v7  |   .1075934   .0194717     5.53   0.000   .0694296   .1457572
      v19 |   .0027636   .000192     14.40   0.000   .0023873   .0031398
  _Iv1_20 | -61.85714    6.665008    -9.28   0.000  -74.92032  -48.79396
[output omitted]
  _Iv1_1292 | -46.3949    3.786434   -12.25   0.000  -53.81617  -38.97363
  _Iv1_1300 | -118.123    10.57277   -11.17   0.000  -138.8453  -97.4008
   _cons |   65.7226    2.600232    25.28   0.000   60.62624   70.81896
-----

```

Let's examine other structures:

```

. xtglm v41 v76 v5 v7 v19, panels(correlated)
panels must be balanced
r(459);

```

```
. xtgls v41 v76 v5 v7 v19, panels(hetero) corr(psar1) force
(note: 2 observations dropped because only 1 obs in group)
```

Cross-sectional time-series FGLS regression

```
Coefficients: generalized least squares
Panels:      heteroskedastic
Correlation: panel-specific AR(1)
```

```
Estimated covariances      =      122      Number of obs      =      4943
Estimated autocorrelations =      122      Number of groups   =      122
Estimated coefficients     =         5      Obs per group: min =         2
                                           avg = 40.51639
                                           max =      148
                                           Wald chi2(4)      =      249.77
                                           Prob > chi2       =      0.0000
```

v41	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
v76	.005669	.0012307	4.61	0.000	.0032568 .0080812
v5	.0151149	.0016583	9.11	0.000	.0118647 .018365
v7	.0431944	.0102922	4.20	0.000	.0230221 .0633668
v19	.0015384	.0001706	9.02	0.000	.0012041 .0018726
_cons	36.84121	3.214902	11.46	0.000	30.54012 43.1423

Article example:

Waldfogel, Jane. 1997. The Effect of Children on Women's Wages. *American Sociological Review*, 62, 209-217.