What are the some of the methods for analyzing clustered data in Stata?

Authored by

stats writer

July 1, 2024

Stata is a statistical software that offers various methods for analyzing clustered data, which is data that is grouped or clustered together in some way. Some of the methods for analyzing clustered data in Stata include cluster-robust standard errors, cluster-robust variance-covariance matrix estimation, and clustered linear regression. These methods take into account the correlation among observations within the same cluster, which can lead to biased results if not properly addressed. Additionally, Stata offers tools such as intra-cluster correlation coefficients and cluster-specific estimates to further explore and understand the clustering effect on the data. These methods are useful in a variety of fields, such as economics, social sciences, and public health, where data is often naturally clustered. Using these methods in Stata allows for more accurate and reliable analysis of clustered data.

What are the some of the methods for analyzing clustered data in Stata? | Stata FAQ

This page was created to show various ways that Stata can analyze clustered data. The intent is to show how the various cluster approaches relate to one another. It is not meant as a way to select a particular model or cluster approach for your data. In selecting a method to be used in analyzing clustered data the user must think carefully about the nature of their data and the assumptions underlying each of the approaches shown below. More examples of analyzing clustered data can be found on our webpage Stata Library: Analyzing Correlated Data.

The dataset we will use to illustrate the various procedures is imm23.dta that was used in the Kreft and de Leeuw
Introduction
to multilevel modeling. This dataset has 519 students clustered in 23 schools. In each of
the models we will regress math on homework. The intraclass correlation for these
data are a hefty 0.30.

We begin by reading in the data.

```
use
https://stats.idre.ucla.edu/stat/examples/imm/imm23,
clear
```

The table below provides a summary of the various models shown that we tried.
Some of the models only adjust the standard error by computing
a cluster robust standard error for the coefficient. Other procedures do more complex
modeling of the multilevel structure. And there are some procedures that do various combinations of

the two.

 # model coef se coef ss residucal bic

1 regress math homework 3.126 .286 48259.9 3837.7

2 regress math homework, cluster(schid) 3.126 .543 48259.9 3837.7

3 svy: regress math homework 3.126 .543 48259.9 **

4 areg math homework, absorb(schid) 2.361 .281 35281.8 3675.1

5 areg math homework, absorb(schid) cluster(schid) 2.361 .650 35281.8 3668.9

6 xtreg math homework, i(schid) fe 2.361 .281 35281.8 3675.1

7 xtreg math homework, i(schid) robust fe 2.361 .321 35281.8 3675.1

8 xtreg math homework, i(schid) re 2.398 .277 35510.8 **

9 xtreg math homework, i(schid) robust re 2.398 .300 35510.8 **

10 xtreg math homework, i(schid) corr(exc) pa 2.383 .259 ?? **

11 xtreg math homework, i(schid) corr(exc) robust pa 2.383 .623 ?? **

12 xtgls math homework, i(schid) panels(iid) 3.126 .286 48259.9 3837.7
13 xtgls math homework, i(schid) panels(hetero) 3.536 .271 48472.9 3805.8

14 xtreg math homework, i(schid) mle 2.402 .277 35283.3 3755.5
15 xtmixed math homework || schid:, mle 2.402 .277 35546.0 3755.5
16 xtmixed math homework || schid:, reml 2.400 .277 35525.0 3754.3

?? population averaged models do not generate residuals with predict
**** bic not avaiable for this procedure**

Below you will find the complete results for each of the above models.

The plain vanilla OLS
that does not account for clustering is included for two reasons: 1) to provide
comparison values for other more appropriate models and 2) because many people
still analyze clustered

data in this manner.

/* model 1 -- plain vanilla OLS */

regress math homework

Source | SS df MS Number of obs = 519
----------+---------- F(1, 517) = 119.43
Model | 11148.1461 1 11148.1461 Prob > F = 0.0000
Residual | 48259.9001 517 93.346035 R-squared = 0.1877
----------+---------- Adj R-squared = 0.1861
Total | 59408.0462 518 114.687348 Root MSE = 9.6616

---

math | Coef. Std. Err. t P>|t|
----------+---

homework | 3.126375 .2860801 10.93 0.000 2.564352 3.688397
_cons | 45.56015 .7055719 64.57 0.000 44.17401 46.94629

/* model 2 -- same as svy: regress with psu */

regress math homework, cluster(schid)

```
Linear regression Number of obs = 519
F( 1, 22) = 33.09
Prob > F = 0.0000
R-squared = 0.1877
Number of clusters (schid) = 23 Root MSE = 9.6616

------------------------------------------------------------------------------
             | Robust
        math | Coef. Std. Err. t P>|t|
-------------+----------------------------------------------------------------
    homework | 3.126375 .5434562 5.75 0.000 1.999315 4.253434
       _cons | 45.56015 1.428639 31.89 0.000 42.59734 48.52297
------------------------------------------------------------------------------

/* model 3 -- same as OLS with cluster option */

svyset schid

pweight:
VCE: linearized
Strata 1:
SU 1: schid
FPC 1: svy: regress math homework
```

(running regress on estimation sample)

Survey: Linear regression

Number of strata = 1 Number of obs = 519
Number of PSUs = 23 Population size = 519
Design df = 22
F(1, 22) = 33.16
Prob > F = 0.0000
R-squared = 0.1877

--
| Linearized
math | Coef. Std. Err. t P>|t|
-------------+--
homework | 3.126375 .5429314 5.76 0.000 2.000404 4.252345
_cons | 45.56015 1.42726 31.92 0.000 42.6002 48.52011
--

/* model 4 -- same as xtreg with fe option */

areg math homework, absorb(schid)

Linear regression, absorbing indicators Number of obs = 519

F(1, 495) = 70.82
Prob > F = 0.0000
R-squared = 0.4061
Adj R-squared = 0.3785
Root MSE = 8.4425

---

math | Coef. Std. Err. t P>|t|
---+---
homework | 2.360971 .2805572 8.42 0.000 1.809741 2.912201
_cons | 47.06884 .6656947 70.71 0.000 45.7609 48.37677
---+---
schid | F(22, 495) = 8.276 0.000 (23 categories)

/* model 5 */

areg math homework, absorb(schid) cluster(schid)

Linear regression, absorbing indicators Number of obs = 519
F(1, 22) = 13.18
Prob > F = 0.0015
R-squared = 0.4061
Adj R-squared = 0.3785

Root MSE = 8.4425

(Std. Err. adjusted for 23 clusters in schid)

| Robust

math | Coef. Std. Err. t P>|t|

---+---

homework | 2.360971 .6502249 3.63 0.001 1.012487 3.709455

_cons | 47.06884 1.281657 36.72 0.000 44.41084 49.72683

---+---

schid | absorbed (23 categories)

/* model 6 -- same as areg */

xtreg math homework, i(schid) fe

Fixed-effects (within) regression Number of obs = 519

Group variable (i): schid Number of groups = 23

R-sq: within = 0.1252 Obs per group: min = 5

between = 0.1578 avg = 22.6

overall = 0.1877 max = 67

F(1,495) = 70.82

corr(u_i, Xb) = 0.2213 Prob > F = 0.0000

---

math | Coef. Std. Err. t P>|t|

---+---

homework | 2.360971 .2805572 8.42 0.000 1.809741 2.912201

_cons | 47.06884 .6656947 70.71 0.000 45.7609 48.37677

---+---

sigma_u | 5.0555127

sigma_e | 8.4425339

rho | .26393678 (fraction of variance due to u_i)

F test that all u_i=0: F(22, 495) = 8.28 Prob > F = 0.0000

/* model 7 */

xtreg math homework, i(schid) robust fe

Fixed-effects (within) regression Number of obs = 519

Group variable (i): schid Number of groups = 23

R-sq: within = 0.1252 Obs per group: min = 5

between = 0.1578 avg = 22.6

overall = 0.1877 max = 67

F(1,495) = 53.95
corr(u_i, Xb) = 0.2213 Prob > F = 0.0000

| Robust
math | Coef. Std. Err. t P>|t|
---+---
homework | 2.360971 .3214347 7.35 0.000 1.729426 2.992516
_cons | 47.06884 .7078415 66.50 0.000 45.67809 48.45958
---+---
sigma_u | 5.0555127
sigma_e | 8.4425339
rho | .26393678 (fraction of variance due to u_i)

/* model 8 */

xtreg math homework, i(schid) re

Random-effects GLS regression Number of obs = 519
Group variable (i): schid Number of groups = 23

R-sq: within = 0.1252 Obs per group: min = 5

between = 0.1578 avg = 22.6
overall = 0.1877 max = 67

Random effects u_i ~ Gaussian Wald chi2(1) = 74.91
corr(u_i, X) = 0 (assumed) Prob > chi2 = 0.0000

---

math | Coef. Std. Err. z P>|z|
---+---
homework | 2.39842 .2771195 8.65 0.000 1.855276 2.941564
_cons | 46.36018 1.17714 39.38 0.000 44.05303 48.66733
---+---
sigma_u | 4.7060369
sigma_e | 8.4425339
rho | .23705881 (fraction of variance due to u_i)

/* model 9 */

xtreg math homework, i(schid) robust re

Random-effects GLS regression Number of obs = 519
Group variable (i): schid Number of groups = 23

R-sq: within = 0.1252 Obs per group: min = 5

between = 0.1578 avg = 22.6
overall = 0.1877 max = 67

Random effects u_i ~ Gaussian Wald chi2(1) = 63.70
corr(u_i, X) = 0 (assumed) Prob > chi2 = 0.0000

| Robust
math | Coef. Std. Err. z P>|z|
---+---
homework | 2.39842 .3004959 7.98 0.000 1.809459 2.987381
_cons | 46.36018 1.185593 39.10 0.000 44.03646 48.6839
---+---
sigma_u | 4.7060369
sigma_e | 8.4425339
rho | .23705881 (fraction of variance due to u_i)

/* model 10 */

xtreg math homework, i(schid) corr(exc) nolog pa

GEE population-averaged model Number of obs = 519
Group variable: schid Number of groups = 23

Link: identity Obs per group: min = 5
Family: Gaussian avg = 22.6
Correlation: exchangeable max = 67
Wald chi2(1) = 84.39
Scale parameter: 94.57586 Prob > chi2 = 0.0000

---

math | Coef. Std. Err. z P>|z|
---+---
homework | 2.382626 .2593715 9.19 0.000 1.874267 2.890985
_cons | 46.41468 1.340197 34.63 0.000 43.78794 49.04142

/* model 11 */

xtreg math homework, i(schid) corr(exc) robust nolog pa

GEE population-averaged model Number of obs = 519
Group variable: schid Number of groups = 23
Link: identity Obs per group: min = 5
Family: Gaussian avg = 22.6
Correlation: exchangeable max = 67

Wald chi2(1) = 14.61
Scale parameter: 94.57586 Prob > chi2 = 0.0001

(Std. Err. adjusted for clustering on schid)

| Semi-robust
math | Coef. Std. Err. z P>|z|

homework | 2.382626 .6232753 3.82 0.000 1.161029 3.604223
_cons | 46.41468 1.632429 28.43 0.000 43.21518 49.61418

/* model 13 -- same as regular OLS */

xtgls math homework, i(schid) panels(iid)

Cross-sectional time-series FGLS regression

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

Estimated covariances = 1 Number of obs = 519
Estimated autocorrelations = 0 Number of groups = 23

Estimated coefficients = 2 Obs per group: min = 5

avg = 22.56522

max = 67

Wald chi2(1) = 119.89

Log likelihood = -1912.6 Prob > chi2 = 0.0000

---

math | Coef. Std. Err. z P>|z|

---+---

homework | 3.126375 .2855283 10.95 0.000 2.566749 3.686

_cons | 45.56015 .7042111 64.70 0.000 44.17992 46.94038

/* model 13 */

xtgls math homework, i(schid) panels(hetero)

Cross-sectional time-series FGLS regression

Coefficients: generalized least squares

Panels: heteroskedastic

Correlation: no autocorrelation

Estimated covariances = 23 Number of obs = 519

Estimated autocorrelations = 0 Number of groups = 23
Estimated coefficients = 2 Obs per group: min = 5
avg = 22.56522
max = 67
Wald chi2(1) = 170.34
Log likelihood = -1896.654 Prob > chi2 = 0.0000

---

math | Coef. Std. Err. z P>|z|
---+---
homework | 3.535692 .2709065 13.05 0.000 3.004725 4.066659
_cons | 44.95865 .6672198 67.38 0.000 43.65092 46.26638

/* model 14 -- same as xtmixed with mle option */

xtreg math homework, i(schid) mle nolog

Random-effects ML regression Number of obs = 519
Group variable (i): schid Number of groups = 23

Random effects u_i ~ Gaussian Obs per group: min = 5
avg = 22.6

max = 67

LR chi2(1) = 70.28
Log likelihood = -1865.247 Prob > chi2 = 0.0000

---

math | Coef. Std. Err. z P>|z|
---+---
homework | 2.401972 .2771764 8.67 0.000 1.858717 2.945228
_cons | 46.34945 1.142005 40.59 0.000 44.11117 48.58774
---+---
/sigma_u | 4.497317 .7861614 3.192697 6.335039
/sigma_e | 8.434685 .2677724 7.925855 8.976181
rho | .2213627 .0616007 .1202136 .3589456

Likelihood-ratio test of sigma_u=0: chibar2(01)= 94.71
Prob>=chibar2 = 0.000

/* model 15 -- same xtreg with mle option */

xtmixed math homework || schid:, mle nolog

Mixed-effects ML regression Number of obs = 519

Group variable: schid Number of groups = 23

Obs per group: min = 5
avg = 22.6
max = 67

Wald chi2(1) = 75.32
Log likelihood = -1865.247 Prob > chi2 = 0.0000

--
math | Coef. Std. Err. z P>|z|
-------------+--
homework | 2.401972 .2767745 8.68 0.000 1.859504 2.94444
_cons | 46.34945 1.141154 40.62 0.000 44.11283 48.58607
--

--
Random-effects Parameters | Estimate Std. Err.
-----------------------------+--
schid: Identity |
sd(_cons) | 4.497318 .7861623 3.192697 6.335042
-----------------------------+--
sd(Residual) | 8.434685 .2677724 7.925855 8.976181

LR test vs. linear regression: chibar2(01) = 94.71 Prob >= chibar2 = 0.0000

/* model 16 */

xtmixed math homework || schid:, nolog

Mixed-effects REML regression Number of obs = 519
Group variable: schid Number of groups = 23

Obs per group: min = 5
avg = 22.6
max = 67

Wald chi2(1) = 74.95
Log restricted-likelihood = -1864.6598 Prob > chi2 = 0.0000

---

math | Coef. Std. Err. z P>|z|
-------------+--
homework | 2.399867 .2771976 8.66 0.000 1.856569 2.943164
_cons | 46.35575 1.162776 39.87 0.000 44.07676 48.63475

Random-effects Parameters | Estimate Std. Err.

---

schid: Identity |

sd(_cons) | 4.619694 .8198156 3.262557 6.541363

---

sd(Residual) | 8.44297 .2682979 7.933157 8.985545

LR test vs. linear regression: chibar2(01) = 96.43 Prob >= chibar2 = 0.0000