The Stata Journal (2022)
22, Number 2, pp. 404–415

DOI: 10.1177/1536867X221106416

Estimating the complier average causal effect via a latent class approach using gsem

Patricio Troncoso
Heriot-Watt University
Edinburgh, U.K.
p.troncoso@hw.ac.uk

Ana Morales-Gómez
University of Edinburgh
Edinburgh, U.K.
Ana.Morales@ed.ac.uk

Abstract. In randomized controlled trials, intention-to-treat analysis is customarily used to estimate the effect of the trial. However, in the presence of noncompliance, this can often lead to biased estimates because intention-to-treat analysis completely ignores varying levels of actual treatment received. This is a known issue that can be overcome by adopting the complier average causal effect approach, which estimates the effect the trial had on the individuals who complied with the protocol. When compliance is unobserved in the control group, the complier average causal effect estimate can be obtained via a latent class specification using the `gsem` command.

Keywords: st0677, gsem, complier average causal effect, randomized control trial, compliance, adherence, latent class modeling, mixture modeling

1 Introduction

In a standard randomized controlled trial (RCT), we compare the outcome of interest in two groups: 1) the treatment group, which is composed of those participants who are randomly selected to a new treatment; and 2) the control group, which is composed of those who are randomly selected to continue with the usual practice (standard treatment or no treatment at all). When all participants who are randomized to treatment receive the treatment in full or comply with all the requirements of the RCT, then the trial arm assignment is enough to estimate the difference between both groups. This is the standard intention-to-treat (ITT) approach.

However, in real-world scenarios, participants assigned to treatment may not receive the treatment in full or engage in all the activities required by the RCT. This is a common occurrence that can bias the ITT estimate, which is why compliance needs to be accounted for.

In this article, we show how to fit a complier average causal effect (CACE) model using the Stata `gsem` command. This type of model can also be fit in Stata by using the community-contributed command `gllamm` (Rabe-Hesketh, Skrondal, and Pickles 2004). Even though CACE models can be fit with `gllamm`, most of `gllamm`'s features have been incorporated in `gsem` from Stata 14 onward.

To the best of our knowledge, there is no worked example available of how to fit a CACE model in Stata using `gsem`. In this article, we explain the main features of this

model and how to fit it, providing a reproducible example and using the main options available in a user-friendly way. We aim to provide a practical introduction to CACE modeling in Stata for researchers who are already familiar with RCT designs, statistical analysis, and Stata's main capabilities at an intermediate level but who may not be acquainted with the latent class approach to CACE and `gsem`'s capabilities.

The next section discusses the CACE approach to addressing the issue of noncompliance in RCTs.

2 Addressing noncompliance in RCTs

The CACE estimate is defined as the difference between the outcome in those participants who complied with the intervention and those participants who would have complied if assigned to treatment. The assumptions of CACE are discussed in detail by Imbens and Rubin (1997) and Little and Yau (1998). Peugh et al. (2017) also provide a thorough introduction to CACE modeling and recent extensions.

The overarching principle of CACE is that the causal effect does not happen because of the mere offer of the treatment itself (ITT) but rather because of the actual treatment received. This is the reasoning behind the "exclusion restriction" assumption in CACE modeling (Little and Yau 1998), which states that the treatment has no effect on the outcome in those participants who did not comply with the intervention.

The main difficulty of estimating the causal effect of compliance is that, while it is straightforward to determine compliance with the intervention in the treatment group, it remains unknown or unobservable in the control group. It is necessary, therefore, to deploy a method that allows distinguishing, within the control group, those who would have complied had they been assigned to treatment from those who would not.

Skrondal and Rabe-Hesketh's (2004) thorough description of the methods to obtain the CACE estimate takes the "latent class with training data" approach proposed by Muthén (2002). This is a probabilistic approach that seeks to estimate the "true" compliance status in the control group while treating observed compliance in the treatment arm as "known".

Consider the example of a school-based intervention where some classrooms are randomized to perform an activity for a prescribed length of time while others are randomized to continue with "business as usual". During the intervention period, researchers record the times at which the activity is conducted in the intervention group and find wide variability. Given that some classrooms performed the activity for longer than others, we might expect the effect of the intervention to be "diluted" by those classrooms that conducted the activity for shorter periods. Dosage would then be key to understand the effect of the trial on the outcome of interest, and this measure can be used to determine compliance status in the intervention group.

Naturally, classrooms under the "business as usual" regime have no records for said activity, but within this group some would have conducted the activity had they had the

chance to do so; hence, "true" compliance in the control group is unknown. Additionally, given that there is heterogeneity in the intervention group (that is, there are compliers and noncompliers), assuming homogeneity in the control group would not necessarily be enabled. By virtue of the randomization itself, the characteristics that make children in the intervention group more likely to comply would also make the children in the control group more likely to comply; that is, there is equivalence of groups prior to intervention (Skrondal and Rabe-Hesketh 2004). This is the rationale behind fitting a CACE model with a latent class approach.

Humphrey et al. (2022) used this approach when analyzing the effect of the "good behavior game" on health- and education-related outcomes in children attending primary schools in England. This RCT aimed to improve classroom behavior through a team game (that is, the treatment) played during school time, and it was expected that this would positively impact on mental health and school attendance mainly.

The following section illustrates how the latent class approach can be implemented in Stata using `gsem`.

2.1 Specifying a CACE model with gsem

Below, we can see the basic specification of a CACE model using `gsem`. This makes use of [SEM] **gsem path notation extensions**. This is a latent class regression model (also known as a mixture model) with specific constraints that are necessary for CACE estimation.

```
gsem ///
(1: depvar <- i.treatment@0 [indepvars] [, family(familyname) ] ) ///
(2: depvar <- i.treatment [indepvars] [, family(familyname) ] ) ///
(C <- [varlist] ) ///
(1: comp <- _cons@-15, logit) ///
(2: comp <- _cons@15, logit), ///
lclass(C 2)
```

The most basic specification of a CACE model would have the following variables:

- *depvar* is the outcome of interest, which is by default assumed to have Gaussian distribution.
- `treatment` is a binary variable to indicate whether the observation was assigned to treatment (`treatment` = 1) or control (`treatment` = 0).
- `comp` is a binary variable to indicate compliance in the treatment group; it is missing in the control group.

The following are optional:

- *indepvars* are the predictors of the outcome of interest.
- *varlist* is the predictor or set of predictors for compliance.

The first line of the syntax is simply the call to `gsem`. The second and third lines compose the regression model for the outcome *depvar*:

`(1:` *depvar* `<- i.treatment@0` [*indepvars*] [`, family(`*familyname*`)`] `)`
`(2:` *depvar* `<- i.treatment` [*indepvars*] [`, family(`*familyname*`)`] `)`

`(1:)` is the regression path for noncompliers, which has the treatment effect fixed to 0; this is the exclusion restriction assumption. `(2:)` is the regression path for compliers, where the treatment effect is estimated freely; this is the CACE estimate. Even though not strictly necessary for the estimation of CACE, it is advisable to fit the model with predictors for the outcome of interest (*indepvars*), for example, baseline measures (Twisk et al. 2018).

The fourth line, `(C)`, is a regression model for the latent class on a set of covariates:

`(C <-` [*varlist*] `)`

This is not strictly essential in the estimation, but it is preferable to have a set of covariates that are reasonably good predictors of compliance (Jo 2002). If no variables are specified or the whole line is omitted, an intercept-only compliance model is fit.

The fifth and sixth lines compose the latent class model for compliance in the treatment group with special constraints that treat compliance in the treatment arm as known:

```
(1: comp <- _cons@-15, logit)
(2: comp <- _cons@15, logit)
```

`_cons@-15` fixes the probability of membership to the compliers class of those who did not comply with the treatment to essentially 0. Meanwhile, `_cons@15` fixes the probability of membership to the compliers class of those who complied with the treatment to essentially 1.

Finally, the last line states the name of the latent class (`C`) and the number of classes, which needs to be set to 2 (compliers and noncompliers):

```
lclass(C 2)
```

Other available options are discussed in section 3.

2.2 An example CACE application

This example uses data from the JOBS II intervention (Vinokur, Price, and Schul 1995). This RCT aimed to prevent depression as a result of job loss by providing training to jobseekers. Our example replicates the CACE estimate as reported by Little and Yau (1998), which is also replicated in Skrondal and Rabe-Hesketh (2004) using `gllamm` (Rabe-Hesketh, Skrondal, and Pickles 2004). The aim here is to estimate the causal effect of actually receiving the treatment, that is, attending job training seminars, on the outcome of interest, that is, depression. A brief description of the variables used in this example is presented in table 1.

Table 1. Variables in the JOBS II dataset (Vinokur, Price, and Schul 1995)

Variable	Description
`depress`	depression score; outcome variable
`depbase`	baseline depression score
`risk`	baseline risk; an index of depression, financial strain, and assertiveness
`r`	dummy variable for being randomized to treatment
`c`	dummy variable for compliance (valid only for treatment group)
`age`	age in years
`motivate`	motivation to attend the job training seminars
`educ`	school grade completed
`assert`	assertiveness
`single`	dummy variable for being single
`econ`	economic hardship
`nonwhite`	dummy for not being white versus being white

First, we read the data from the `gllamm` website, as such:

```
. infile depress risk r depbase age motivate educ
> assert single econ nonwhite x10 c c0
> using "http://www.gllamm.org/books/wjobs.dat", clear
(502 observations read)
```

The variable `c` is a dummy to indicate compliance in the treatment group. From this variable, we need to derive the variable `comp`, which is missing (unobserved) in the control group, as such:

```
. generate comp=c if r==1
(167 missing values generated)
```

Then, we specify and run the CACE model. We constrained the effects of covariates in the regression equations to be equal across classes, that is, `@c1` and `@c2`. These additional constraints were specified to replicate the results in Little and Yau (1998) but are not essential for other applications.

```
. gsem (1.C: depress <- i.r@0 depbase@c1 risk@c2)
> (2.C: depress <- i.r depbase@c1 risk@c2)
> (C <-  age motivate educ assert
> single econ nonwhite)
> (1:comp <- _cons@-15, logit)
> (2:comp <- _cons@15, logit),
> lclass(C 2) nolog
Generalized structural equation model                     Number of obs = 502
Log likelihood = -729.41415
 ( 1)  [comp]1bn.C = -15
 ( 2)  [depress]1.r#1bn.C = 0
 ( 3)  [depress]1bn.C#c.depbase - [depress]2.C#c.depbase = 0
 ( 4)  [depress]1bn.C#c.risk - [depress]2.C#c.risk = 0
 ( 5)  [comp]2.C = 15
 ( 6)  [/]var(e.depress)#1bn.C - [/]var(e.depress)#2.C = 0
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.C	(base outcome)					
2.C						
age	.0790447	.0140223	5.64	0.000	.0515615	.1065279
motivate	.6668729	.159823	4.17	0.000	.3536257	.9801202
educ	.2997692	.0675169	4.44	0.000	.1674386	.4320999
assert	-.3758715	.146405	-2.57	0.010	-.6628201	-.088923
single	.5401949	.2754367	1.96	0.050	.0003489	1.080041
econ	-.1586017	.1596608	-0.99	0.321	-.4715312	.1543278
nonwhite	-.4985881	.3123487	-1.60	0.110	-1.11078	.1136041
_cons	-8.740022	1.581572	-5.53	0.000	-11.83985	-5.640198

```
Class:    1
Response: depress                                         Number of obs = 502
Family:   Gaussian
Link:     Identity
Response: comp                                            Number of obs = 335
Family:   Bernoulli
Link:     Logit
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
depress						
1.r	0	(omitted)				
depbase	-1.463379	.1826867	-8.01	0.000	-1.821438	-1.10532
risk	.9117568	.2624529	3.47	0.001	.3973586	1.426155
_cons	1.632537	.2791255	5.85	0.000	1.085461	2.179613
comp						
_cons	-15	(constrained)				
var(e.depress)	.506397	.0322776			.4469262	.5737814

```
Class:    2
Response: depress                                     Number of obs = 502
Family:   Gaussian
Link:     Identity
Response: comp                                        Number of obs = 335
Family:   Bernoulli
Link:     Logit
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
depress						
1.r	-.3098673	.1173219	-2.64	0.008	-.5398141	-.0799205
depbase	-1.463379	.1826867	-8.01	0.000	-1.821438	-1.10532
risk	.9117568	.2624529	3.47	0.001	.3973586	1.426155
_cons	1.81249	.2971227	6.10	0.000	1.23014	2.394839
comp						
_cons	15	(constrained)				
var(e.depress)	.506397	.0322776			.4469262	.5737814

The CACE estimate is given on the table for the regression in class 2. The effect of treatment (`r`) in the compliers class is -0.3098673, which indicates that those who complied in the intervention arm are expected to score 0.31 less in the depression scale than those who would have complied if assigned to treatment.

With `gsem`, many postestimation functions are available. For example, we can request a summary of model fit information (useful for comparing competing models) by typing

```
. estat ic
Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	502	.	-729.4141	14	1486.828	1545.889

```
Note: BIC uses N = number of observations. See [R] BIC note.
```

Next we illustrate how we can use `margins` and `marginsplot` to obtain predicted values per latent class. First, we run the `margins` command to obtain some predicted values across classes at the quartiles of baseline depression (`depbase`) in the noncompliers class (class 1):

```
. margins, at((p25) depbase) at((p50) depbase) at((p75) depbase)
> predict(outcome(depress) class(1))
Predictive margins                                      Number of obs = 502
Model VCE: OIM

Expression: Predicted mean (depress in class 1.C), predict(outcome(depress)
            class(1))
1._at: depbase = 2.27 (p25)
2._at: depbase = 2.45 (p50)
3._at: depbase = 2.64 (p75)
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	-.1575271	.0621197	-2.54	0.011	-.2792795	-.0357748
2	-.4209354	.0542381	-7.76	0.000	-.5272402	-.3146306
3	-.6989775	.0657265	-10.63	0.000	-.8277991	-.5701559

Second, we run the `marginsplot` command to visualize the predicted values in the noncompliers class (class 1):

```
. marginsplot, title("Noncompliers (overall)") xtitle("Predicted depression")
> ytitle("Baseline depression") name(class1, replace)
> recast(scatter)
> ylabel(1 "2.27" 2 "2.45" 3 "2.64")
> xlabel(-1(.2).5)
> plotopts(msymbol(Oh))
> horizontal xline(0, lpattern(dash))
> scheme(sj)
Variables that uniquely identify margins: _atopt
Multiple at() options specified:
      _atoption=1: (p25) depbase
      _atoption=2: (p50) depbase
      _atoption=3: (p75) depbase
  (output omitted)
```

We omitted the plot because this is an intermediate step for the final plot presented in figure 1. We also requested the predicted values for the noncompliers class overall. There is no need to do this by treatment group—its predicted values are constrained to equality because we specified the exclusion restriction assumption.

Third, we run the `margins` command again but this time for the predicted values in the compliers class (class 2) across treatment groups:

```
. margins r, at((p25) depbase) at((p50) depbase) at((p75) depbase)
> predict(outcome(depress) class(2))
Predictive margins                                      Number of obs = 502
Model VCE: OIM

Expression: Predicted mean (depress in class 2.C), predict(outcome(depress)
            class(2))
1._at: depbase = 2.27 (p25)
2._at: depbase = 2.45 (p50)
3._at: depbase = 2.64 (p75)
```

	Delta-method					
	Margin	std. err.	z	P>|z|	[95% conf.	interval]
_at#r						
1 0	.0224253	.1128368	0.20	0.842	-.1987308	.2435813
1 1	-.287442	.0601775	-4.78	0.000	-.4053878	-.1694963
2 0	-.2409831	.1043744	-2.31	0.021	-.4455532	-.036413
2 1	-.5508504	.0527557	-10.44	0.000	-.6542496	-.4474512
3 0	-.5190252	.1063014	-4.88	0.000	-.727372	-.3106783
3 1	-.8288925	.065151	-12.72	0.000	-.956586	-.7011989

Afterward, we call the `marginsplot` command to visualize the predicted values in the compliers class (class 2) by trial arm (plot omitted).

```
. marginsplot, plotdimension(r, labels("Control" "Treatment"))
> title("Compliers") xtitle("Predicted depression")
> ytitle("Baseline depression") name(class2, replace)
> ylabel(1 "2.27" 2 "2.45" 3 "2.64")
> xlabel(-1(.2).5)
> plot1opts(msymbol(Sh))
> plot2opts(msymbol(S))
> recast(scatter)
> horizontal xline(0, lpattern(dash))
> legend(size(*0.8) position(0) bplacement(neast) rows(2))
> scheme(sj)
Variables that uniquely identify margins: r _atopt
Multiple at() options specified:
      _atoption=1: (p25) depbase
      _atoption=2: (p50) depbase
      _atoption=3: (p75) depbase
```

(*output omitted*)

Finally, we call `graph combine` to plot the predicted values by complier status and trial arm (figure 1):

```
. graph combine class1 class2, cols(2) scheme(sj)
```

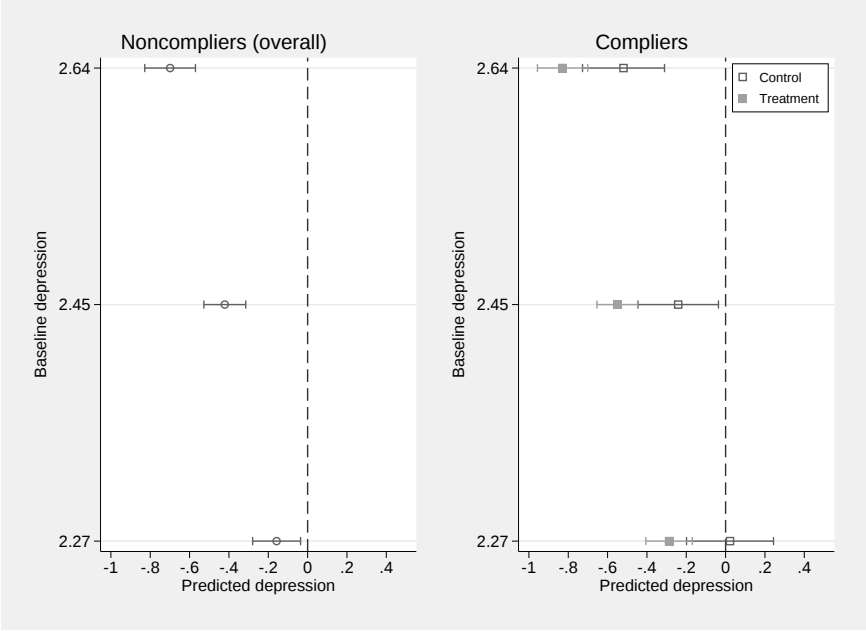

Figure 1. Predicted scores by complier status and trial arm

In figure 1, we can see that compliers (right-hand side) in the control group have higher predicted scores for depression at all values of baseline depression. On the other hand, noncompliers (regardless of trial arm) have predicted depression scores lower than compliers in the control group but not as low as compliers in the treatment group. The same steps can be repeated for "risk" scores at baseline to visualize the effect of compliance at different levels of risk.

Some additional options are discussed in the next section.

3 Conclusions

In this article, we presented the main features of a CACE model using `gsem` in Stata. This was aimed at researchers already familiar with RCT data analysis and Stata itself, using a reproducible and well-known example. Interested readers are encouraged to explore further options available with `gsem`, and we mention some potentially useful ones next.

`gsem` comes with many postestimation functions that can be used in CACE modeling. In section 2.2, we provided an example to obtain the information criteria, which can be

used for model comparison and selection, as well as an example of margins and visualization. For more details of the available functions, see [SEM] **gsem postestimation**.

In addition, equality constraints can be used potentially to test other hypotheses of interest, such as common or separate coefficients across compliers and noncompliers. For more details, see [SEM] **sem and gsem option constraints( )** and [SEM] **sem and gsem path notation**.

With `gsem`, one can also fit CACE models for a variety of distributions, hence the option of specifying `family()` in the above general form of the syntax. For more details of the allowed distribution families and link functions, see [SEM] **gsem family-and-link options**.

Finally, clustering around groups can be accounted for with the `vce()` option. This is especially relevant for cluster RCTs, where whole groups go through the process of randomization before the intervention. Humphrey et al. (2022) recently applied this approach using `gsem` to fit a CACE model in a school-based RCT in England. This study fit CACE models for generalized outcomes using clustered standard errors, providing a further example of `gsem`'s functions and extended capabilities. These CACE estimates (obtained using `gsem`) are presented in Troncoso (2021).

To sum up, `gsem` is a flexible command that can be used to tackle a variety of statistical problems pertaining to the analysis of trial efficacy, including under noncompliance as shown here. The versatility of `gsem` and Stata's overall usability makes it a powerful tool for researchers analyzing RCT data.

4 Acknowledgments

This work was supported by the Economic and Social Research Council through the following grants: 1) Understanding Children's Lives and Outcomes and 2) Scottish Centre for Administrative Data Research (grant numbers ES/V011243/1 and ES/S007407/1, respectively). This work was also supported in its early stages by the National Institute for Health Research (grant number 14/52/38). The authors are affiliated with the Scottish Centre for Administrative Data Research (SCADR), which is part of the Administrative Data Research U.K. partnership, funded by the Economic and Social Research Council.

5 References

Humphrey, N., A. Hennessey, P. Troncoso, M. Panayiotou, L. Black, K. Petersen, L. Wo, and et al. 2022. The Good Behaviour Game intervention to improve behavioural and other outcomes for children aged 7–8 years: A cluster RCT. Unpublished manuscript.

Imbens, G. W., and D. B. Rubin. 1997. Estimating outcome distributions for compliers in instrumental variable models. *Review of Economic Studies* 64: 555–574. https://doi.org/10.2307/2971731.

Jo, B. 2002. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Statistics in Medicine* 21: 3161–3181. https://doi.org/10.1002/sim.1267.

Little, R. J., and L. H. Y. Yau. 1998. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods* 3: 147–159. https://doi.org/10.1037/1082-989X.3.2.147.

Muthén, B. O. 2002. Beyond SEM: General latent variable modeling. *Behaviormetrika* 29: 81–117. https://doi.org/10.2333/bhmk.29.81.

Peugh, J. L., D. Strotman, M. McGrady, J. Rausch, and S. Kashikar-Zuck. 2017. Beyond intent to treat (ITT): A complier average causal effect (CACE) estimation primer. *Journal of School Psychology* 60: 7–24. https://doi.org/10.1016/j.jsp.2015.12.006.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2004. GLLAMM manual. Working Paper 160, Division of Biostatistics, University of California–Berkeley. https://biostats.bepress.com/ucbbiostat/paper160/.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Troncoso, P. 2021. A "CACE" in point: Estimating causal effects via a latent class approach in RCTs with noncompliance using Stata. Presented August 5–6, 2021, at the Stata Conference 2021. https://www.stata.com/meeting/us21/slides/US21_Troncoso.pdf.

Twisk, J., L. Bosman, T. Hoekstra, J. Rijnhart, M. Welten, and M. Heymans. 2018. Different ways to estimate treatment effects in randomised controlled trials. *Contemporary Clinical Trials Communications* 10: 80–85. https://doi.org/10.1016/j.conctc.2018.03.008.

Vinokur, A. D., R. H. Price, and Y. Schul. 1995. Impact of the JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology* 23: 39–74. https://doi.org/10.1007/BF02506922.

About the authors

Patricio Troncoso is an applied statistician and a research fellow with the SCADR at Heriot-Watt University and an Honorary Research Fellow at the University of Manchester. His research focuses on a range of educational and children's outcomes, such as attainment, school value-added, and child protection.

Ana Morales-Gómez is an applied statistician and a research fellow with the SCADR at the University of Edinburgh. Her research uses quantitative methods for understanding inequalities with an emphasis on crime and justice.