xthdidlogit — Heterogeneous difference-in-differences panel-data logistic regression[+]

[+] This command is part of StataNow.

Description | Quick start | Menu | Syntax
Options | Remarks and examples | Stored results | Methods and formulas
Acknowledgments | References | Also see

Description

`xthdidlogit` estimates average treatment effects on the treated (ATETs) that may vary over time and over treatment cohorts for binary outcomes using pooled logistic regression.

Quick start

Estimate ATETs of treatment `treat` on binary outcome `y` with group `grpvar`, and model `y` with covariate `x` on `xtset` data

```
xthdidlogit (y x) (treat), group(grpvar)
```

Same as above, but use user-provided cohort variable `mycohort` instead of letting the command generate a cohort variable

```
xthdidlogit (y x) (treat), group(grpvar) usercohort(mycohort)
```

Same as above, but use the not-yet-treated group as the control group

```
xthdidlogit (y x) (treat), group(grpvar) controlgroup(notyet) ///
    usercohort(mycohort)
```

Same as above, but without including interactions between covariates and the treatment or interactions between covariates and the cohort variable

```
xthdidlogit (y x) (treat), group(grpvar) controlgroup(notyet) ///
    usercohort(mycohort) noxinteract
```

Menu

Statistics > Causal inference/treatment effects > Binary outcomes > Panel-data heterogeneous DID

Syntax

```
xthdidlogit (ovar [omvarlist]) (tvar) [if] [in] [weight], group(groupvar) [options]
```

ovar is a binary outcome of interest.

omvarlist specifies the covariates in the outcome model and may contain factor variables; see **[U] 11.4.3 Factor variables**.

tvar must be a binary variable indicating observations subject to treatment.

groupvar is a categorical variable that indicates the group level at which the treatment occurs.

options	Description
Model	
* group(*groupvar*)	specify group variable
controlgroup(*cgtype*)	specify the type of control group; default is controlgroup(never)
cohortvar(*cvar*[, replace])	specify the name for the generated cohort variable
usercohort(*varname*)	specify the name of the cohort variable to be used for estimation
hettype(*hetspec*)	specify the type of heterogeneity; default is hettype(timecohort)
over(*overspec*)	specify the subpopulation over which each ATET is computed; default is over(cohort)
noxinteract	omit covariate interactions
SE/Robust	
vce(*vcetype*)	*vcetype* may be cluster *clustvar*, robust, bootstrap, or jackknife
Reporting	
level(#)	set confidence level; default is level(95)
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
coeflegend	display legend instead of statistics

cgtype	Description
never	use the never-treated group as the control group; the default
notyet	use the not-yet-treated group as the control group

hetspec	Description
timecohort	heterogeneous treatment effects over time and cohort; the default
time	heterogeneous treatment effects over time
cohort	heterogeneous treatment effects over cohort

overspec	Description
`cohort`	use cohort subpopulations when controlling for covariate heterogeneity; the default
`timecohort`	use time and cohort subpopulations when controlling for covariate heterogeneity

*`group()` is required.

A panel variable and a time variable must be specified using `xtset`; see [XT] **xtset**.

`by`, `collect`, and `statsby` are allowed; see **[U] 11.1.10 Prefix commands**.

`fweights`, `iweights`, and `pweights` are allowed; see **[U] 11.1.6 weight**. Weights must be constant within panel.

`coeflegend` does not appear in the dialog box.

See **[U] 20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

Model

`group(`*groupvar*`)` specifies a group variable that indicates the group level at which the treatment occurs. *groupvar* may be, for example, states, counties, or hospitals. `group()` also defines the clusters for the default cluster–robust standard errors. `group()` is required. You may specify only one group variable.

`controlgroup(`*cgtype*`)` specifies the type of control group. A control group can be either a never-treated group or a not-yet-treated group. A never-treated group refers to the units that are untreated from the first to the last period. A not-yet-treated group refers to the units that are untreated up to a specific period. *cgtype* can be one of `never`, referring to the never-treated group, or `notyet`, referring to the not-yet-treated group. The default is `controlgroup(never)`.

`cohortvar(`*cvar*[`, replace`]`)` specifies the variable name *cvar* for the generated cohort variable. The cohort variable is a categorical variable indicating the period when the unit is first treated. By default, `_did_cohort` is used as the name of the cohort variable. If `_did_cohort` already exists in the dataset, it is replaced if the `cohortvar()` option is not specified.

If the `replace` suboption is specified, *cvar* is replaced.

`usercohort(`*varname*`)` specifies a variable to be used as a cohort indicator during estimation. By default, a cohort variable is generated using the information in the estimation sample to indicate the period when a unit is first treated. `usercohort()` overrides this default and allows you to provide a cohort indicator. This is useful, for instance, when there are gaps in the estimation sample, but you know a group was treated at the time when the gap is present in the data.

`hettype(`*hetspec*`)` specifies time or cohort heterogeneity for the `twfe` estimator. By default, treatment is interacted with time and cohort. You may choose to keep one of time or cohort interactions using *hetspec*.

hetspec may be one of `timecohort` for heterogeneous treatment effects over both time and cohort, `time` for heterogeneous treatment effects over time only, or `cohort` for heterogeneous treatment effects over cohort only. The default is `hettype(timecohort)`.

`over(`*overspec*`)` specifies the subpopulation over which treatment effects are averaged. When covariates are specified, you need to control for their variation when computing treatment effects. The default is `over(cohort)`, meaning effects are averaged over each cohort and adjusted for the time period effect. If there are no covariates, this option has no effect.

When the heterogeneity type is specified as `timecohort` (`hettype(timecohort)`), *overspec* may be one of `cohort` for effects averaged over each cohort or `timecohort` for effects averaged over each time period and cohort.

When the heterogeneity type is specified as `cohort` (`hettype(cohort)`), *overspec* may only be `cohort` for effects averaged over each cohort.

`over()` may not be specified with `hettype(time)`.

`noxinteract` specifies that the covariates in *omvarlist* not be interacted with the difference-in-differences regressors. Covariates are still included in the model as levels.

SE/Robust

`vce(`*vcetype*`)` specifies the type of standard error reported, which includes types that allow for intragroup correlation (`cluster` *clustvar*), that are robust to intragroup correlation among group variable (`robust`), and that use bootstrap or jackknife sampling done at the panel level (`bootstrap`, `jackknife`); see [R] ***vce_option***.

`vce(cluster` *clustvar*`)`, the default, uses the variable specified in `group(`*groupvar*`)`.

Specifying `vce(robust)` is equivalent to specifying `vce(cluster` *clustvar*`)`, where *clustvar* is the variable specified in the `group(`*groupvar*`)` option.

Reporting

`level(#)`; see [R] **Estimation options**.

display_options: `noci`, `nopvalues`, `cformat(%`*fmt*`)`, `pformat(%`*fmt*`)`, `sformat(%`*fmt*`)`, and `nolstretch`; see [R] **Estimation options**.

The following option is available with `xthdidlogit` but is not shown in the dialog box:

`coeflegend`; see [R] **Estimation options**.

Remarks and examples

It is common to study the effects of a treatment, for example, a policy or intervention, on a group. `xthdidlogit` is for data where the treated groups are subject to the treatment at different points in time and they remain exposed to the treatment. Outcomes are either zero or nonzero as with `logit`.

`xthdidlogit` estimates ATET parameters that change over time and treatment cohorts (groups treated at different points in time). Each one of these ATETs has the same interpretation that the parameters of a two-time, two-group difference-in-differences (DID) parameter would have. Because there are multiple DID parameters, we refer to them as heterogeneous treatment effects or as heterogeneous DID. This contrasts with estimating only one ATET, which assumes there is no variation across time or cohort. If you assume no variation across time or cohort and you do not want to fit the model by using logistic regression, you may use `xtdidregress`; see [CAUSAL] **didregress**.

`xthdidlogit` implements a pooled logistic regression estimator described in Wooldridge (2023). It fits a cross-sectional model in which the treatment is interacted with time, cohort, and covariates. This approach allows you to incorporate treatment heterogeneity at all the aforementioned levels. Covariate interactions allow you to explore moderating effects (how the treatment effect is affected by covariates) by using the `estat moderation` postestimation command. Cohort and time interactions allow

you to explore heterogeneity within time and cohort by using `estat aggregation, time` and `estat aggregation, cohort`. You may also want to explore how treatment effects evolve by using `estat aggregation, dynamic`.

For treatment effects to be identified in a DID framework, a parallel-trends assumption needs to be satisfied; see [CAUSAL] **DID intro**. In the linear case, fit using `xthdidregress twfe`, the parallel-trends assumption is satisfied for the linear prediction, $\mathbf{x}\beta$, of the potential outcomes. The linear prediction is, by construction, a one-to-one mapping of the conditional mean in that case. For nonlinear DID, you can think about the parallel-trends assumption as being satisfied either for conditional means or for the linear prediction but not for both. For the estimator implemented in `xthdidlogit`, the identification relies on parallel trends for the linear prediction as proposed by Wooldridge (2023). You may use `estat ptrends` to evaluate this assumption.

Example 1: Fitting a logistic heterogeneous DID model

We are interested in modeling the effect of having no social media apps, `nosmapps`, on high levels of stress, `hstress`, for a panel of high school students. `hstress` is based on a diagnostic test and is 0 if the student does not experience a high level of stress and 1 otherwise. The school district promoted eliminating social media apps from phones starting in 2014. Throughout the time of the study, some students opted to follow the suggestion, while others did not. As controls, we incorporate the hours exercised each week, `exercise`, and whether the student participates in a group activity, `gactivity`. Group activities could be sports or after-school activities or clubs, none of which are necessarily sponsored by the school.

To obtain the effect of `nosmapps` on `hstress`, we first `xtset` our panel dataset:

```
. use https://www.stata-press.com/data/r19/hstress
(Fictional data on high stress levels and social media usage)
. xtset id year
Panel variable: id (unbalanced)
 Time variable: year, 2011 to 2016, but with gaps
         Delta: 1 unit
```

We fit the model with the `xthdidlogit` command. In the first set of parentheses, we declare the outcome and any covariates that affect the outcome directly. In the second set of parentheses, we define the observation-level treatment variable. After the comma, we need to define the group variable in `group()`; this is a required option. The group variable defines at which level the treatment occurs and also identifies the clustering variable, which in this case is `id`. We also use the `over()` option with argument `timecohort`. This is advisable if you have time-varying controls in your panel-data specification. Our command and results are as follows:

```
. xthdidlogit (hstress exercise i.gactivity) (nosmapps), group(id)
> over(timecohort)
note: variable _did_cohort, containing cohort indicators formed by treatment
      variable nosmapps and group variable id, was added to the dataset using
      the estimation sample.
Computing ATETs using margins ...
Treatment and time information
Time variable: year
Time interval: 2011 to 2016
Control:       _did_cohort = 0
Treatment:     _did_cohort > 0
```

	_did_cohort
Number of cohorts	4
Number of obs	
Never treated	2151
2014	790
2015	190
2016	106

```
Heterogeneous-treatment-effects logistic regression   Number of obs    = 3,237
                                                      Number of panels =   545
Panel variable:  id
Treatment level: id
Control group:   Never treated
Heterogeneity:   Cohort and time
                                  (Std. err. adjusted for 545 clusters in id)
```

Cohort	ATET	Robust std. err.	z	P>|z|	[95% conf.	interval]
2014						
year						
2014	-.0907947	.0314546	-2.89	0.004	-.1524445	-.0291449
2015	-.0888262	.0315567	-2.81	0.005	-.1506761	-.0269763
2016	-.0877322	.0295599	-2.97	0.003	-.1456687	-.0297958
2015						
year						
2015	-.0887931	.0459465	-1.93	0.053	-.1788465	.0012604
2016	-.0197798	.0472443	-0.42	0.675	-.1123768	.0728173
2016						
year						
2016	-.0261036	.094283	-0.28	0.782	-.2108948	.1586876

```
Note: ATET computed using covariates.
```

Notice the note below the command. A variable with the name _did_cohort has been generated. Looking at the group variable and the observation-level treatment, xthdidlogit detected we have treatment-time cohorts. The new variable creates treatment groups based on the time when a group was first treated. For instance, if a teenager deleted social media apps from their phone in 2014, they were grouped into the 2014 cohort. The variable also contains a category for a control group. In this case, the control group is formed by the teenagers that, during the period of analysis, never deleted their social media apps.

Cohorts are an important input for estimation and for postestimation commands. We do not need to adhere to the default name, `_did_cohort`, and may provide our own name by using the `cohortvar()` option. Also, we do not need to rely on `xthdidlogit` creating the cohort based on the estimation sample; instead, we might provide a cohort indicator we created or which is already in our dataset by using the `usercohort()` option.

Next appears a table that gives us a sense of the treatment groups and time. We see the time variable, `year`, and its range, `2011 to 2016`. Then we see what defines a treated or a control group. The table below provides group-level information about the cohort-time groups. The first row tells us the number of cohorts. Following the number of cohorts is a tabulation showing how many observations are in each cohort. For instance, 2,151 observations are never treated in our dataset. The table gives us a sense of the amount of information available in each cohort and might hint at the variability of cohort-level estimates.

The next table presents the ATET estimates. The first panel shows the ATETs for the 2014 cohort. We first have the 2014 ATET of −0.091, with a confidence interval that does not include 0. We should interpret the ATET to mean that, among the students who deleted social media apps from their phones starting in 2014, the probability of having high stress in 2014 is 0.091 lower than if they had not deleted their social media apps. In 2015, we observe a treatment effect on the treated for `hstress` of −0.089, and, on the last year of our data for the 2014 cohort, we see a treatment effect of −0.088. We interpret the results for the remaining cohorts similarly.

◁

▷ Example 2: Looking at pretrends

When we fit the model by using `logit`, as detailed in *Methods and formulas*, we imposed the no anticipation and parallel-trends assumption. We did this by not including interactions between cohorts, time, and treatment for the preintervention periods for each cohort. We can relax these assumptions by incorporating the preintervention interactions and then testing that the treatment effects for these preintervention periods are jointly zero. This pretrend test is given by

```
. estat ptrends
Parallel-trends test (pretreatment time period)
H0: Treatment effects in all the pretreatment periods are zero
chi2(  9)   = 9.10
Prob > chi2 = 0.4280
```

We cannot reject the null hypothesis that the preintervention effects are zero.

◁

Stored results

xthdidlogit stores the following in e():

Scalars

e(N)	number of observations
e(N_clust)	number of clusters
e(N_panels)	number of panels
e(tmin)	first time period
e(tmax)	last time period
e(rank)	rank of e(V)
e(xinteract)	1 if covariate interactions are included, 0 otherwise

Macros

e(cmd)	xthdidlogit
e(cmdline)	command as typed
e(depvar)	name of outcome variable
e(clustvar)	name of cluster variable
e(control_group)	control group
e(het_type)	heterogeneity type
e(over_type)	subpopulation over which each ATET is computed
e(cohortvar)	name of cohort variable
e(usercohort)	name of user-specified cohort variable
e(ovar)	name of outcome variable
e(wtype)	weight type
e(wexp)	weight expression
e(marginsnotok)	predictions disallowed by margins
e(ivar)	panel identifier variable
e(timevar)	time variable
e(treatname)	name of treatment variable
e(estat_cmd)	program used to implement estat
e(vce)	*vcetype* specified in vce()
e(vcetype)	title used to label Std. err.
e(properties)	b V

Matrices

e(b)	coefficient vector
e(V)	variance–covariance matrix of the estimators
e(cohort_count)	matrix with cohort count information

Functions

e(sample)	marks estimation sample

In addition to the above, the following is stored in r():

Matrices

r(table)	matrix containing the coefficients with their standard errors, test statistics, *p*-values, and confidence intervals

Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

Methods and formulas

Methods and formulas are presented under the following headings:

The model
The pooled logistic regression estimator

The model

Denote $y_{it}(g)$ as the potential outcome for unit i at time t if it was first treated at time g. Denote $y_{it}(0)$ as the potential outcome for unit i at time t if it is never treated. All the units can be grouped into cohorts, and the units of each cohort start the treatment simultaneously. For example, cohort g means all the units start treatment at time g. Let G_{ig} be an indicator that equals 1 if unit i is first treated at time g. Then the units in cohort g can be denoted by $G_{ig} = 1$. When a unit i is never treated, we denote $G_{i0} = 1$. Thus, cohort 0 indicates all the units that are never treated. We assume that once a unit is treated, it will remain treated.

To reveal how the heterogeneous treatment effects evolve across cohorts and time, we are interested in estimating the ATET for each combination of cohort and time. Let $\theta(g,t)$ be the ATET for cohort g at time t, which is defined as

$$\theta(g,t) = \mathbf{E}\left\{y_t(g) - y_t(0) | G_g = 1\right\} \tag{ATET}$$

where G_g equals 1 if a unit belongs to cohort g. We use `margins` to estimate $\theta(g,t)$ in equation (ATET). We cannot directly estimate $\theta(g,t)$ by using equation (ATET) because the potential outcomes $y_t(g)$ and $y_t(0)$ are not observable.

The pooled logistic regression estimator

We estimate the parameters of the model by using pooled logistic regression; see [R] **logit**. The model can be thought of in terms of the conditional probability of a positive outcome or, what is equivalent, the mean of the outcome conditional of the covariates, which is given by

$$\begin{aligned} E\left(y_{it}|\mathbf{x}_i, d_{it}, G_{ig}, f_s\right) = \Lambda\Big(&\eta + \sum_{g=q}^{T} G_{ig}\alpha_g + \sum_{s=q}^{T} f_s\gamma_s + \sum_{g=q}^{T}\sum_{s=g}^{T} d_{it}G_{ig}f_s\tau_{gs} + \sum_{g=q}^{T} G_{ig}\mathbf{x}_i\boldsymbol{\alpha}_g^{\mathbf{x}} \\ &+ \sum_{s=q}^{T} f_s\mathbf{x}_i\boldsymbol{\gamma}_s^{\mathbf{x}} \sum_{g=q}^{T}\sum_{s=g}^{T} d_{it}G_{ig}f_s\mathbf{x}_i\boldsymbol{\tau}_{gs}^{\mathbf{x}} + \mathbf{x}_{it}\boldsymbol{\beta}\Big) \end{aligned} \tag{LGT}$$

In the expression above, Λ is the cumulative distribution function of the logistic distribution. The terms on the right-hand side of equation (LGT) are the regressors in the model for which the cohort-time treatment effects are a function of τ_{gs}, $\boldsymbol{\tau}_{gs}^{\mathbf{x}}$, and the covariates. In equation (LGT), d_{it} is a treatment indicator, and f_s is an indicator for posttreatment periods with s going from q to T, where q is the first time period we observe treatment. For instance, f_q equals 1 if we are in time period q and 0 otherwise. We can compute these effects using `margins` by typing

```
. margins, dydx(d) at(year=q ... year=T) over(cohort) vce(unconditional)
```

where `d` is the treatment indicator, `year` indicates treatment times at which treatment will be evaluated using `at()`, and `cohort` is the treatment-time cohorts.

If `over(timecohort)` was specified with `xthdidlogit`, we can obtain effects by typing

```
. margins, dydx(d) over(year cohort) vce(unconditional)
```

We use `vce(unconditional)` to account for the variation in the covariates. When there are no covariates specified or when `noxinteract` is specified, `vce(unconditional)` has no effect. Furthermore, if we include covariates but specify `noxinteract`, our equation reduces to

$$E\left(y_{it}|\mathbf{x}_{it},d_{it},G_{ig},f_s\right)=\Lambda\Big(\eta+\sum_{g=q}^{T}G_{ig}\alpha_g+\sum_{s=q}^{T}f_s\gamma_s+\sum_{g=q}^{T}\sum_{s=g}^{T}d_{it}G_{ig}f_s\tau_{gs}+\mathbf{x}_{it}\boldsymbol{\beta}\Big)$$

With the `hettype()` option, we reduce the complexity of equation (LGT). In particular, if we ask for `hettype(time)`, we have

$$\begin{aligned}E\left(y_{it}|\mathbf{x}_{it},d_{it},G_{ig},f_s\right)=\Lambda\Big(&\eta+\sum_{g=q}^{T}G_{ig}\alpha_g+\sum_{s=q}^{T}f_s\gamma_s+\sum_{s=q}^{T}d_{it}f_s\tau_s+\sum_{g=q}^{T}G_{ig}\mathbf{x}_i\boldsymbol{\alpha}_g^{\mathbf{x}}\\&+\sum_{s=q}^{T}f_s\mathbf{x}_i\boldsymbol{\gamma}_s^{\mathbf{x}}+\sum_{s=q}^{T}d_{it}f_s\mathbf{x}_i\boldsymbol{\tau}_s^{\mathbf{x}}+\mathbf{x}_{it}\boldsymbol{\beta}\Big)\end{aligned}$$

Now treatment varies over time but not over cohort. If we use the `hettype(cohort)` option, we have

$$\begin{aligned}E\left(y_{it}|\mathbf{x}_{it},d_{it},G_{ig},f_s\right)=\Lambda\Big(&\eta+\sum_{g=q}^{T}G_{ig}\alpha_g+\sum_{s=q}^{T}f_s\gamma_s+\sum_{g=q}^{T}d_{it}G_{ig}\tau_g+\sum_{g=q}^{T}G_{ig}\mathbf{x}_i\boldsymbol{\alpha}_g^{\mathbf{x}}\\&+\sum_{s=q}^{T}f_s\mathbf{x}_i\boldsymbol{\gamma}_s^{\mathbf{x}}+\sum_{g=q}^{T}d_{it}G_{ig}\mathbf{x}_i\boldsymbol{\tau}_g^{\mathbf{x}}+\mathbf{x}_{it}\boldsymbol{\beta}\Big)\end{aligned}$$

Now treatment varies over cohort but not over time.

When the `controlgroup(notyet)` option is specified, the G_{ig} indicator excludes the last treated cohort. As discussed in Wooldridge (2025), when every group is eventually treated, we cannot identify the treatment effect for this cohort. It is therefore sensible to use the last treated cohort as a control group. When some of the units in our sample are never treated, we can always identify all cohorts, and the pooled logistic regression estimator will always revert to using `controlgroup(never)`.

Acknowledgments

We thank Fernando Rios-Avila of the London School of Economics and Political Science for helpful conversations and advice and for his contributions in developing heterogeneous treatment-effects software and accompanying material. His community-contributed command `jwdid` implements some of the methods in this entry and a few others.

We thank Jeff Wooldridge of Michigan State University for helpful conversations regarding flexible estimation of nonlinear DID models for panel data.

References

Wooldridge, J. M. 2023. Simple approaches to nonlinear difference-in-differences with panel data. *Econometrics Journal* 26: C31–C66. https://doi.org/10.1093/ectj/utad016.

———. 2025. Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. *Empirical Economics* 69: 2545–2587. https://doi.org/10.1007/s00181-025-02807-z.

Also see

[CAUSAL] **xthdidlogit postestimation** — Postestimation tools for xthdidlogit+

[CAUSAL] **DID intro** — Introduction to difference-in-differences estimation

[CAUSAL] **gencohort** — Create a cohort variable

[CAUSAL] **teffects intro** — Introduction to treatment effects for observational data

[CAUSAL] **teffects intro advanced** — Advanced introduction to treatment effects for observational data

[R] **logit** — Logistic regression, reporting coefficients

[U] 20 Estimation and postestimation commands

Stata, Stata Press, Mata, NetCourse, and NetCourseNow are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow is a trademark of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.

For suggested citations, see the FAQ on citing Stata documentation.