areg — Linear regression with a large dummy-variable set

**Description**

areg fits a linear regression absorbing one categorical factor. areg is designed for datasets with many groups, but not a number of groups that increases with the sample size. See the xtreg, fe command in [XT] xtreg for an estimator that handles the case in which the number of groups increases with the sample size.

**Quick start**

Linear regression of y on x, absorbing an indicator variable for each level of cvar

```stata
areg y x, absorb(cvar)
```

As above, but add categorical variable a

```stata
areg y x i.a, absorb(cvar)
```

With cluster–robust standard errors

```stata
areg y x i.a, absorb(cvar) vce(cluster cvar2)
```

Using svyset data

```stata
svy: areg y x i.a, absorb(cvar)
```

**Menu**

Statistics > Linear models and related > Other > Linear regression absorbing one cat. variable
Syntax

areg depvar [ indepvars ] [ if ] [ in ] [ weight ] , absorb(varname) [ options ]

<table>
<thead>
<tr>
<th>options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>*absorb(varname) specifies the categorical variable, which is to be included in the regression as if it were specified by dummy variables. absorb() is required.</td>
</tr>
<tr>
<td>SE/Robust</td>
<td>vce(vcetype) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (ols), that are robust to some kinds of misspecification (robust), that allow for intragroup correlation (cluster clustvar), and that use bootstrap or jackknife methods (bootstrap, jackknife); see [R] vce_option.</td>
</tr>
<tr>
<td>Reporting</td>
<td>level(#) set confidence level; default is level(95)</td>
</tr>
<tr>
<td></td>
<td>display_options control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling</td>
</tr>
<tr>
<td></td>
<td>coeflegend display legend instead of statistics</td>
</tr>
</tbody>
</table>

*absorb(varname) is required.

indepvars may contain factor variables; see [U] 11.4.3 Factor variables.
depvar and indepvars may contain time-series operators; see [U] 11.4.4 Time-series varlists.
bootstrap, by, fp, jackknife, mi estimate, rolling, and statsby are allowed; see [U] 11.1.10 Prefix commands.
vce(bootstrap) and vce(jackknife) are not allowed with the mi estimate prefix; see [MI] mi estimate.
Weights are not allowed with the bootstrap prefix; see [R] bootstrap.
aweights are not allowed with the jackknife prefix; see [R] jackknife.
aweights, fweights, and pweights are allowed; see [U] 11.1.6 weight.
coeflegend does not appear in the dialog box.
See [U] 20 Estimation and postestimation commands for more capabilities of estimation commands.

Options
areg — Linear regression with a large dummy-variable set

Reporting

\[ \text{level(#)} ; \text{see [R] Estimation options.} \]
\[ \text{display_options: noci, nopvalues, nomitted, vsquish, noomitted, baselevels, allbaselevels, nolabel, fwrapp(#), fwrappon(style), cformat(\%fmt), pformat(\%fmt), sformat(\%fmt), and nolstretch; see [R] Estimation options.} \]

The following option is available with areg but is not shown in the dialog box:
\[ \text{coeflegend; see [R] Estimation options.} \]

### Remarks and examples

Suppose that you have a regression model that includes among the explanatory variables a large number, \( k \), of mutually exclusive and exhaustive dummies:
\[ y = X\beta + d_1\gamma_1 + d_2\gamma_2 + \cdots + d_k\gamma_k + \epsilon \]

For instance, the dummy variables, \( d_i \), might indicate countries in the world or states of the United States. One solution would be to fit the model with `regress`, but this solution is possible only if \( k \) is small enough so that the total number of variables (the number of columns of \( X \) plus the number of \( d_i \)'s plus one for \( y \)) is sufficiently small—meaning that it can fit in a Stata matrix (see [R] Limits). If you have more variables than can fit in a Stata matrix, `regress` will not work. `areg` provides a way of obtaining estimates of \( \beta \)—but not the \( \gamma_i \)'s—in these cases. The effects of the dummy variables are said to be absorbed.

#### Example 1

So that we can compare the results produced by `areg` with Stata’s other regression commands, we will fit a model in which \( k \) is small. `areg`’s real use, however, is when \( k \) is large.

In our automobile data, we have a variable called `rep78` that is coded 1, 2, 3, 4, and 5, where 1 means poor and 5 means excellent. Let’s assume that we wish to fit a regression of `mpg` on `weight`, `gear_ratio`, and `rep78` (parameterized as a set of dummies).
. use https://www.stata-press.com/data/r16/auto2
(1978 Automobile Data)
. regress mpg weight gear_ratio b5.rep78
Source | SS df MS Number of obs = 69
---------+-------------------------------------------------------------
Model | 1575.97621 6 262.662702 Prob > F = 0.0000
Residual | 764.226686 62 12.3262369 R-squared = 0.6734
Total | 2340.2029 68 34.4147485 Adj R-squared = 0.6418
---------+-------------------------------------------------------------
mpg | Coef. Std. Err. t P>|t| [95% Conf. Interval]
---------+-------------------------------------------------------------
weight | -.0051031 .0009206 -5.54 0.000 -.0069433 -.003263
gear_ratio | .901478 1.565552 0.58 0.567 -2.228015 4.030971
rep78 | Poor | -2.036937 2.740728 -0.74 0.460 -7.515574 3.4417
Fair | -2.419822 1.764338 -1.37 0.175 -5.946682 1.107039
Average | -2.557432 1.370912 -1.87 0.067 -5.297846 .1829814
Good | -2.788389 1.395259 -2.00 0.050 -5.577473 .0006939
_cons | 36.23782 7.01057 5.17 0.000 22.22389 50.25175
---------+-------------------------------------------------------------

To fit the areg equivalent, we type

. areg mpg weight gear_ratio, absorb(rep78)
Linear regression, absorbing indicators
Absorbed variable: rep78
Number of obs = 69
No. of categories = 5
F( 2, 62) = 41.64
Prob > F = 0.0000
R-squared = 0.6734
Adj R-squared = 0.6418
Root MSE = 3.5109

mpg | Coef. Std. Err. t P>|t| [95% Conf. Interval]
---------+-------------------------------------------------------------
weight | -.0051031 .0009206 -5.54 0.000 -.0069433 -.003263
gear_ratio | .901478 1.565552 0.58 0.567 -2.228015 4.030971
_cons | 34.05889 7.056383 4.83 0.000 19.95338 48.1644
---------+-------------------------------------------------------------

F test of absorbed indicators: F(4, 62) = 1.117     Prob > F = 0.356

Both regress and areg display the same $R^2$ values, root mean squared error, and—for weight and gear_ratio—the same parameter estimates, standard errors, $t$ statistics, significance levels, and confidence intervals. areg, however, does not report the coefficients for rep78, and, in fact, they are not even calculated. This computational trick makes the problem manageable when $k$ is large. areg reports a test that the coefficients associated with rep78 are jointly zero. Here this test has a significance level of 35.6%. This $F$ test for rep78 is the same that we would obtain after regress if we were to specify test 1.rep78 2.rep78 3.rep78 4.rep78; see [R] test.

The model $F$ tests reported by regress and areg also differ. The regress command reports a test that all coefficients except that of the constant are equal to zero; thus, the dummies are included in this test. The areg output shows a test that all coefficients excluding the dummies and the constant are equal to zero. This is the same test that can be obtained after regress by typing test weight gear_ratio.
Technical note

`areg` is designed for datasets with many groups, but not a number that grows with the sample size. Consider two different samples from the U.S. population. In the first sample, we have 10,000 individuals and we want to include an indicator for each of the 50 states, whereas in the second sample we have 3 observations on each of 10,000 individuals and we want to include an indicator for each individual. `areg` was designed for datasets similar to the first sample in which we have a fixed number of groups, the 50 states. In the second sample, the number of groups, which is the number of individuals, grows as we include more individuals in the sample. For an estimator designed to handle the case in which the number of groups grows with the sample size, see the `xtreg, fe` command in [XT] `xtreg`.

Although the point estimates produced by `areg` and `xtreg, fe` are the same, the estimated VCEs differ when `vce(cluster clustvar)` is specified because the commands make different assumptions about whether the number of groups increases with the sample size.

Technical note

The intercept reported by `areg` deserves some explanation because, given $k$ mutually exclusive and exhaustive dummies, it is arbitrary. `areg` identifies the model by choosing the intercept that makes the prediction calculated at the means of the independent variables equal to the mean of the dependent variable: $\hat{y} = \bar{x} \hat{\beta}$.

```
. predict yhat
   (option xb assumed; fitted values)
. summarize mpg yhat if rep78 != .
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>69</td>
<td>21.28986</td>
<td>5.866408</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>yhat</td>
<td>69</td>
<td>21.28986</td>
<td>4.383224</td>
<td>11.58643</td>
<td>28.07367</td>
</tr>
</tbody>
</table>

We had to include `if rep78 < .` in our `summarize` command because we have missing values in our data. `areg` automatically dropped those missing values (as it should) in forming the estimates, but `predict` with the `xb` option will make predictions for cases with missing `rep78` because it does not know that `rep78` is really part of our model.

These predicted values do not include the absorbed effects (that is, the $d_i \gamma_i$). For predicted values that include these effects, use the `xb`d option of `predict` (see [R] `areg postestimation`) or see [XT] `xtreg`.

Example 2

`areg, vce(robust)` is a Huberized version of `areg`; see [P] `__robust`. Just as `areg` is equivalent to using `regress` with dummies, `areg, vce(robust)` is equivalent to using `regress, vce(robust)` with dummies. You can use `areg, vce(robust)` when you expect heteroskedastic or nonnormal errors. `areg, vce(robust)`, like ordinary regression, assumes that the observations are independent, unless the `vce(cluster clustvar)` option is specified. If the `vce(cluster clustvar)` option is specified, this independence assumption is relaxed and only the clusters identified by equal values of `clustvar` are assumed to be independent.
Assume that we were to collect data by randomly sampling 10,000 doctors (from 100 hospitals) and then sampling 10 patients of each doctor, yielding a total dataset of 100,000 patients in a cluster sample. If in some regression we wished to include effects of the hospitals to which the doctors belonged, we would want to include a dummy variable for each hospital, adding 100 variables to our model. \texttt{areg} could fit this model by

\[
\texttt{. areg depvar patient\_vars, absorb(hospital) vce(cluster doctor)}
\]

\section*{Stored results}
\texttt{areg} stores the following in \texttt{e()}: 

\begin{itemize}
  \item Scalars 
    \begin{itemize}
      \item \texttt{e(N)} \quad \text{number of observations}
      \item \texttt{e(k\_absorb)} \quad \text{number of absorbed categories}
      \item \texttt{e(mss)} \quad \text{model sum of squares}
      \item \texttt{e(tss)} \quad \text{total sum of squares}
      \item \texttt{e(df\_m)} \quad \text{model degrees of freedom}
      \item \texttt{e(rss)} \quad \text{residual sum of squares}
      \item \texttt{e(df\_r)} \quad \text{residual degrees of freedom}
      \item \texttt{e(r2)} \quad \text{R}^2
      \item \texttt{e(r2\_a)} \quad \text{adjusted R}^2
      \item \texttt{e(df\_a)} \quad \text{degrees of freedom for absorbed effect}
      \item \texttt{e(rmse)} \quad \text{root mean squared error}
      \item \texttt{e(ll)} \quad \text{log likelihood}
      \item \texttt{e(ll\_0)} \quad \text{log likelihood, constant-only model}
      \item \texttt{e(N\_clust)} \quad \text{number of clusters}
      \item \texttt{e(F)} \quad \text{F statistic}
      \item \texttt{e(F\_absorb)} \quad \text{F statistic for absorbed effect (when vce(robust) is not specified)}
      \item \texttt{e(p)} \quad \text{p-value for model F test}
      \item \texttt{e(p\_absorb)} \quad \text{p-value for F test of absorbed effect}
      \item \texttt{e(rank)} \quad \text{rank of e(V)}
    \end{itemize}

  \item Macros 
    \begin{itemize}
      \item \texttt{e(cmd)} \quad \texttt{areg}
      \item \texttt{e(cmdline)} \quad \text{command as typed}
      \item \texttt{e(depvar)} \quad \text{name of dependent variable}
      \item \texttt{e(absvar)} \quad \text{name of absorb variable}
      \item \texttt{e(wtype)} \quad \text{weight type}
      \item \texttt{e(wexp)} \quad \text{weight expression}
      \item \texttt{e(title)} \quad \text{title in estimation output}
      \item \texttt{e(clustvar)} \quad \text{name of cluster variable}
      \item \texttt{e(vcetype)} \quad \text{\texttt{vcetypen} \ specified in \texttt{vce()}}
      \item \texttt{e(datasignature)} \quad \text{the checksum}
      \item \texttt{e(datasignaturevars)} \quad \text{variables used in calculation of checksum}
      \item \texttt{e(properties) b V}
      \item \texttt{e(predict)} \quad \text{program used to implement \texttt{predict}}
      \item \texttt{e(footnote)} \quad \text{program used to implement the footnote display}
      \item \texttt{e(marginsnotok)} \quad \text{predictions disallowed by margins}
      \item \texttt{e(asbalanced)} \quad \text{factor variables \texttt{fvset} as \texttt{asbalanced}}
      \item \texttt{e(asobserved)} \quad \text{factor variables \texttt{fvset} as \texttt{asobserved}}
    \end{itemize}

  \item Matrices 
    \begin{itemize}
      \item \texttt{e(b)} \quad \text{coefficient vector}
      \item \texttt{e(V)} \quad \text{variance–covariance matrix of the estimators}
      \item \texttt{e(V\_modelbased)} \quad \text{model-based variance}
    \end{itemize}

  \item Functions 
    \begin{itemize}
      \item \texttt{e(sample)} \quad \text{marks estimation sample}
    \end{itemize}
\end{itemize}
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

\texttt{areg} begins by recalculating \textit{depvar} and \textit{indepvars} to have mean 0 within the groups specified by \texttt{absorb()}. The overall mean of each variable is then added back in. The adjusted \textit{depvar} is then regressed on the adjusted \textit{indepvars} with \texttt{regress}, yielding the coefficient estimates. The degrees of freedom of the variance–covariance matrix of the coefficients is then adjusted to account for the absorbed variables—this calculation yields the same results (up to numerical roundoff error) as if the matrix had been calculated directly by the formulas given in \([R]\) \texttt{regress}.

\texttt{areg} with \texttt{vce(robust)} or \texttt{vce(cluster clustvar)} works similarly, calling \_\texttt{robust} after \texttt{regress} to produce the Huber/White/sandwich estimator of the variance or its clustered version. See \([P]\) \_\texttt{robust}, particularly \textit{Introduction} and \textit{Methods and formulas}. The model \( F \) test uses the robust variance estimates. There is, however, no simple computational means of obtaining a robust test of the absorbed dummies; thus, this test is not displayed when the \texttt{vce(robust)} or \texttt{vce(cluster clustvar)} option is specified.

The number of groups specified in \texttt{absorb()} are included in the degrees of freedom used in the finite-sample adjustment of the cluster–robust VCE estimator. This statement is valid only if the number of groups is small relative to the sample size. (Technically, the number of groups must remain fixed as the sample size grows.) For an estimator that allows the number of groups to grow with the sample size, see the \texttt{xtreg}, fe command in \([XT]\) \texttt{xtreg}.

### References


### Also see

\[R\] \texttt{areg postestimation} — Postestimation tools for \texttt{areg}

\[R\] \texttt{regress} — Linear regression

\[MI\] \texttt{Estimation} — Estimation commands for use with \texttt{mi estimate}

\[XT\] \texttt{xtreg} — Fixed-, between-, and random-effects and population-averaged linear models

\[U\] \textit{20 Estimation and postestimation commands}