

Censored Least Absolute Deviations Estimator: CLAD

July 12, 2000

Dean Jolliffe, Center for Economic Research and Graduate Education, Czech Republic,
dean.jolliffe@cerge.cuni.cz

Bohdan Krushelnytskyy, Center for Economic Research and Graduate Education, Czech
Republic, bohdan.krushelnytskyy@cerge.cuni.cz

Anastassia Semykina, Center for Economic Research and Graduate Education, Czech
Republic, anastassia.semykina@cerge.cuni.cz

Description

This insert provides a program for estimating Powell's censored least absolute deviations estimator (CLAD) and bootstrap estimates of its sampling variance. The CLAD estimator is a generalization of the least absolute deviations (LAD) estimator, which is implemented in Stata by *qreg*. Unlike the standard estimators of the censored regression model such as Tobit or other maximum likelihood approaches, the CLAD estimator is robust to heteroscedasticity and is consistent and asymptotically normal for a wide class of error distributions. (See Arabmazar and Schmidt, 1981, and Vijverberg, 1987, for empirical examples of the magnitude of the bias resulting from the Tobit in the presence of non-normal error distributions.)

This program sidesteps the issue of programming analytical standard errors and provides instead, bootstrapped estimates of the sampling variance. Rogers (1993) shows that the standard errors reported by Stata for *qreg* are not robust to violations of homoscedasticity or independence of the residuals and proposes a bootstrap alternative. We follow Rogers for the CLAD estimator and propose two bootstrap estimates of the standard errors. The first is the standard bootstrap which assumes that the sample was selected using a simple random design. The second is a bootstrap estimate which assumes that the sample was selected in two-stages, and which replicates the design by bootstrapping in two stages.

An advantage of the two-stage bootstrap estimates available in *clad* is that if the sample was collected using a two-stage process, then the estimated standard errors will be robust to this design effect. Kish (1995) and Cochran (1997) show the importance of correcting mean values for design effects. Scott and Holt (1982) show that the magnitude of the bias for the estimated variance-covariance matrix for ordinary least squares estimates can be quite large when it is erroneously assumed that the data were collected using a simple random sample if in fact a two-stage design had been used.

Syntax

```
clad varlist [if exp] [in range] [, reps(#) psu(varname) ll(#)] ul(#)] dots saving(filename)  
replace level(#) quantile(#) iterate(#) wlsiter(#)]
```

Options

reps(#) specifies the number of bootstrap replications to be performed. The default value is 100.

psu(*varname*) specifies the variable identifying the primary sampling unit. If no variable is specified, then the bootstrap replication is a single-stage, simple random draw on the sample.

ll(#) and **ul**(#) are as in *Stata's tobit* command and indicate the censoring point. **ll**() indicates left censoring and **ul**() indicates right-censoring. If **ll** or **ul** is specified without a specific censoring value, then *clad* assumes that the lower limit is the minimum observed in the data (if **ll** is specified) and the upper limit is the maximum (if **ul** is specified). If nothing is specified for a lower or upper bound, *clad* assumes that the lower limit is zero. *Clad* only functions with lower *or* upper censoring, you can not specify censoring at both the lower and upper bound.

dots prints a dot to the screen for each bootstrap replication thereby allowing the user to estimate, after a few replications, the time to completion.

saving(filename) creates a Stata data file (.dta file) containing the bootstrap sample of the parameter estimates.

replace overwrites the Stata data file specified in **saving**(), if it already exists.

All other options are as specified in *Stata's qreg* command.

Examples

To illustrate the use of *clad*, we use data from the 1988 Ghana Living Standard Survey (GLSS) and consider a somewhat nonsensical regression. The sample considered is 1,581 households and the dependent variable, *loffinc*, is the log of household, non-farm income. Since some households are fully engaged in farming, this variable has 528 observations with zeros recorded. This variable is regressed on the log of the size of the household, *lsize*, and two geographic dummy variables, *urban* and *coastal*. When we issue the *clad* command, the following results:

```
clad loffinc lsize urban coastal, ll(0) reps(200)
```

```
Initial sample size = 1581
Final sample size = 1580
Pseudo R2 = .05048178
```

```
Bootstrap statistics
```

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]		
lsize	200	1.149846	.0555843	.2313103	.6937123	1.60598	(N)
					.8176686	1.712571	(P)
					.8173234	1.699596	(BC)
urban	200	2.375166	-.0195581	.3350349	1.714492	3.03584	(N)
					1.622379	2.967376	(P)
					1.629229	2.981997	(BC)
coastal	200	1.287741	.0120871	.3076209	.6811259	1.894356	(N)
					.7033063	1.974158	(P)
					.697911	1.957236	(BC)
const	200	6.443694	-.0570418	.5362573	5.386218	7.50117	(N)
					5.196927	7.403755	(P)
					5.310441	7.479608	(BC)

N = normal, P = percentile, BC = bias-corrected

The first line of output tells us that the original sample size is 1581 and in the second line we learn that the algorithm for estimation dropped one case from the sample. An important caveat to the Pseudo R-squared reported on the third line, is that this is the reported statistic from the last iteration of the `qreg` command on the final sample size. It is not the Pseudo R-squared for original sample, but we have opted to report this statistic to provide some indication of how the model is performing.

In the example above, no sample design information is passed to `clad` and the program calls Stata's `bsample` utility to re-sample the data 200 times. In order to maintain the same sample size in each bootstrap re-sample, `clad` ignores observations where the dependent variable is missing. The results from `bsample` are then passed to the `bstat` command to generate the standard Stata bootstrap output. For more information about the normal, percentile, and bias-corrected percentile confidence intervals, see `bstrap` in the Stata manuals. For an introduction to the bootstrap principle, see Efron and Tibshirani (1993). In order to reproduce results from `clad` it is necessary to set the random number seed first. (See `generate` in the Stata reference manuals for more information.)

The reported standard errors above will be correct if the sample comes from a simple random draw. This is not the case with the GLSS data, which was collected using a two-stage design. `Clad` can generate bootstrap estimates of the standard errors which are robust to the two-stage design by passing the information about the primary sampling unit to `clad`. So, for example, we correct the standard errors above for this aspect of the sample design by issuing the following command:

```
clad loffinc lsize urban coastal, ll(0) reps(200) psu(clust)
```

```
Initial sample size = 1581
Final sample size = 1580
Pseudo R2 = .05048178
```

```
Bootstrap statistics
```

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]		
lsize	200	1.149846	.1624274	.4902858	.1830239	2.116669	(N)
					.551284	2.321528	(P)
					.4481114	2.026139	(BC)
urban	200	2.375166	.0909652	.7013571	.9921202	3.758212	(N)
					1.233815	4.06944	(P)
					1.162497	3.970161	(BC)
coastal	200	1.287741	.0300567	.580344	.1433277	2.432154	(N)
					.2134603	2.442016	(P)
					.1055352	2.31571	(BC)
const	200	6.443694	-.3422176	1.276675	3.926146	8.961242	(N)
					3.41127	8.008909	(P)
					4.029336	8.207627	(BC)

N = normal, P = percentile, BC = bias-corrected

It is worth noting that introducing information about the sample design only affects the estimates of the standard errors. The dramatic increase in the size of the standard errors is not that surprising as the design effect for the dependent variable is approximately 3.8 and there is little in the observation matrix which will explain the intra-cluster correlation.

Methods and Formulas

Powell's (1984) CLAD estimator is found by minimizing

$$\sum |y_i - \max(0, x_i' \beta)| \quad (1)$$

The consistency of this estimator rests on the fact that medians are preserved by monotone transformations of the data, and (1) is a monotone transformation of the standard least absolute deviations (LAD) regression. The properties of the LAD estimator are presented in Koenker and Basset (1978). The LAD estimator is implemented in Stata with the **qreg** command.

The estimation technique used in this *ado* for the CLAD estimator is Buchinsky's (1994) iterative linear programming algorithm (ILPA). (For a critique of and alternative to this algorithm, see Fitzenberger, 1997.) The first step of the ILPA is to estimate a quantile regression for the full sample, then delete the observations for which the predicted value of the dependent variable is less than zero. Another quantile regression is estimated on the new sample, and again negative predicted values are dropped. More generally, observations are dropped if the predicted value is less than the censoring value when the left tail of the distribution is censored, or they are dropped if the predicted value is greater than the censoring value when the right tail of the distribution is censored. Buchinsky (1991) shows that if the process converges, then a local minimum is obtained. Convergence occurs when there are no negative predicted values in two consecutive iterations.

The two bootstraps are implemented as follows. For the simple random sample (srs) we simply use Stata's *bsample* utility to bootstrap the CLAD point estimates. The srs, two-stage bootstrap follows this process: In the first stage it counts the number of unique PSUs, say k , and then using Stata's *uniform* function, randomly selects with replacement k (not necessarily unique) PSUs. At this point it counts the number of times each PSU has been selected and this is stored for later use. To implement the second stage, the program first counts the number of USUs, say m , in each selected PSU and then randomly selects m USUs from each selected PSU. If a PSU is selected more than once, say α times, then in the second stage the program randomly selects αm USUs from the selected PSU. As a final note, we warn that this *ado* can be quite time consuming since the entire algorithm described above is repeated for each bootstrap re-sampling of the data.

References

Arabmazar, A. and Schmidt, P. 1981. "Further Evidence on the Robustness of the Tobit Estimator to Heteroskedasticity," *Journal of Econometrics*, 17: 253-258.

- Buchinsky, M. 1991. "Methodological Issues in Quantile Regression," Chapter 1 of *The Theory and Practice of Quantile Regression*, Ph.D. dissertation, Harvard University.
- Buchinsky, M. 1994. "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica*, 62(2): 405-459.
- Cochran, William, *Sampling Techniques*, New York: Wiley, Third Edition, 1997.
- Efron, Bradley and Robert Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, 57, New York: Chapman Hall, 1993.
- Fitzenberger, Bernd, 1997. "Computational Aspects of Censored Quantile Regression," In Dodge, Y. ed., *Proceedings of The 3rd International Conference on Statistical Data Analysis based on the L_1 - Norm and Related Methods*, pp. 171-186. Institute of Mathematical Statistics Lecture Notes – Monograph Series, Volume 31, Hayward, California.
- Rogers, W., "Calculation of Quantile Regression Standard Errors," *Stata Technical Bulletin*, 1993, STB-13: 18-19.
- Kish, Leslie, *Survey Sampling*, New York: Wiley Classics Library Edition, 1995.
- Koenker, R. and Bassett, G. 1978. "Regression Quantiles." *Econometrica* 46(1): 33-50.
- Powell, J.L. 1984. "Least Absolute Deviations Estimation for the Censored Regression Model." *Journal of Econometrics*, 25: 303-325.
- Scott, A.J. and Holt, D. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." *Journal of American Statistical Association* 77(380): 848-854.
- Vijverberg, W. 1987. "Non-Normality as Distributional Misspecification in Single-Equation Limited Dependent Variable Models." *Oxford Bulletin of Economics and Statistics*. 49(4): 417-430.