The Journal of Open Source Software

LinRegOutliers: A Julia package for detecting outliers in linear regression

Mehmet Hakan Satman[1], **Shreesh Adiga**[2], **Guillermo Angeris**[3], **and Emre Akadal**[4]

1 Department of Econometrics, Istanbul University, Istanbul, Turkey **2** Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru, India **3** Department of Electrical Engineering, Stanford University, Stanford, California, USA **4** Department of Informatics, Istanbul University, Istanbul, Turkey

DOI: 10.21105/joss.02892

Software

- Review
- Repository
- Archive

Editor: Mikkel Meyer Andersen

Reviewers:

- @salleuska
- @rMassimiliano

Submitted: 02 December 2020
Published: 05 January 2021

Summary

`LinRegOutliers` is a Julia package that implements a number of outlier detection algorithms for linear regression. The package also implements robust covariance matrix estimation and graphing functions which can be used to visualize the regression residuals and distances between observations, with many possible metrics (*e.g.*, the Euclidean or Mahalanobis distances with either given or estimated covariance matrices). Our package implements many algorithms and diagnostics for model fitting with outliers under a single interface, which allows users to quickly try many different methods with reasonable default settings, while also providing a good starting framework for researchers who may want to extend the package with novel methods.

State of the field

In linear regression, we are given a number of data points (say, n) where each data point is represented by a vector x_i, with p entries, and a dependent variable that corresponds to each of these data points, represented by the scalar y_i, for $i = 1, 2, \ldots, n$. We then seek to find a linear model which best describes the data (up to some error term, ϵ_i):

$$y_i = \beta_1 (x_i)_1 + \cdots + \beta_p (x_i)_p + \epsilon_i,$$

where $\beta_1, \ldots, \beta_p$ are the p unknown parameters. We will assume that the ϵ_i are independent and identically-distributed (i.i.d.) error terms with zero mean. Note that, if $(x_i)_1 = 1$ for all $i = 1, \ldots, n$, this is equivalent to having an intercept term given by β_1.

We can write this more conveniently by letting X be the *design matrix* of size $n \times p$, whose ith row is given by the vectors x_i (where $(x_i)_1 = 1$ if the model has an intercept), while y is an n-vector of observations, whose entries are y_i, and similarly for ϵ:

$$y = X\beta + \epsilon.$$

The usual approach to finding an estimate for β, which we call $\hat{\beta}$, is the Ordinary Least Squares (OLS) estimator given by $\hat{\beta} = (X^T X)^{-1} X^T y$, which is efficient and has good statistical properties when the error terms are all of roughly the same magnitude (*i.e.*, there are no outliers). On the other hand, the OLS estimator is very sensitive to outliers: even if a

single observation lies far from the regression hyperplane, OLS will often fail to find a good estimate for the parameters, β.

To solve this problem, a number of methods have been developed in the literature. These methods can be roughly placed in one or more of the five following categories: diagnostics, direct methods, robust methods, multivariate methods, and visual methods. *Diagnostics* are methods which attempt to find points that significantly affect the fit of a model (often, such points can be labeled as outliers). Diagnostics can then be used to initialize *direct methods*, which fit a (usually non-robust) model to a subset of points suspected to be clear of outliers; remaining points which are not outliers with respect to this fit are continually added to this subset until all points not in the subset are deemed outliers. *Robust methods*, on the other hand, find a best-fit model by approximately minimizing a loss function that is resistant to outliers. Some of the proposed methods are also *multivariate methods*, which can accommodate obtaining robust location and scale measures of multivariate data. *Visual methods* generally work on the principle of visualizing the statistics obtained from these mentioned methods. As an example, the method `mveltsplot` constructs a 2-D plot using robust distances and scaled residuals obtained from `mve` and `lts` which are multivariate data and robust regression methods, respectively. Many direct and robust methods for regression select an initial basic or clean subset of observations using the results of diagnostics and methods for multivariate data. This is why methods that are not directly related to regression are included in the package.

Statement of need

In practice, many of the proposed methods have reasonable performance and yield similar results for most datasets, but sometimes differ widely in specific circumstances by means of masking and swamping ratios. Additionally, some of the methods are relatively complicated and, if canonical implementations are available, they are often out of date or only found in specific languages of the author's choice, making it difficult for researchers to compare the performance of these algorithms on their datasets.

We have reimplemented many of the algorithms available in the literature in Julia (Bezanson et al., 2017), an open-source, high performance programming language designed primarily for scientific computing. Our package, `LinRegOutliers`, is a comprehensive and simple-to-use Julia package that includes many of the algorithms in the literature for detecting outliers in linear regression. The implemented `Julia` methods for diagnostics, direct methods, robust methods, multivariate methods, and visual diagnostics are shown in **Table 1**, **Table 2**, **Table 3**, **Table 4**, and **Table 5**, respectively.

Table 1: Regression Diagnostics

Algorithm	Reference	Method
Hadi Measure	(Chatterjee & Hadi, 2015)	`hadimeasure`
Covariance Ratio	(Belsley et al., 2005)	`covratio`
DFBETA	(Belsley et al., 2005)	`dfbeta`
DFFIT	(Belsley et al., 2005)	`dffit`
Mahalanobis Distances	(Mahalanobis, 1930)	`mahalanobisSquaredMatrix`
Cook Distances	(Cook, 1977)	`cooks`

Table 2: Direct Methods

Algorithm	Reference	Method
Ransac	(Fischler & Bolles, 1987)	`ransac`

Algorithm	Reference	Method
KS-89	(Kianifard & Swallow, 1989)	`ks89`
HS-93	(Hadi & Simonoff, 1993)	`hs93`
Atkinson-94	(Atkinson, 1994)	`atkinson94`
PY-95	(Peña & Yohai, 1995)	`py95`
SMR-98	(Sebert et al., 1998)	`smr98`
ASM-2000	(Adnan et al., 2000)	`asm2000`
BACON	(Billor et al., 2000)	`bacon`
Imon-2005	(Imon, 2005)	`imon2005`
bch	(Billor et al., 2006)	`bch`

Table 3: Robust Methods

Algorithm	Reference	Method
Least Absolute Deviations	(Nobakhti et al., 2009)	`lad`
Least Absolute Trimmed Deviations	(Hawkins & Olive, 1999)	`lta`
Least Median of Squares	(Rousseeuw, 1984)	`lms`
Least Trimmed Squares	(Rousseeuw & Van Driessen, 2000)	`lts`
CM-97	(Chatterjee & Mächler, 1997)	`cm97`
ga-lts	(Satman, 2012)	`galts`
Satman-2013	(Satman, 2013)	`satman2013`
Satman-2015	(Satman, 2015)	`satman2015`
CCF	(Barratt et al., 2020)	`ccf`

Table 4: Multivariate Methods

Algorithm	Reference	Method
Hadi-1992	(Hadi, 1992)	`hadi1992`
Hadi-1994	(Hadi, 1994)	`hadi1994`
Minimum Volume Ellipsoid	(Van Aelst & Rousseeuw, 2009)	`mve`
Minimum Covariance Determinant	(Rousseeuw & Driessen, 1999)	`mcd`

Table 5: Visual Methods

Algorithm	Reference	Method
BCH Plot	(Billor et al., 2006)	`bchplot`
MVE-LTS Plot	(Van Aelst & Rousseeuw, 2009)	`mveltsplot`
Data Images	(Marchette & Solka, 2003)	`dataimage`
Stalactite Plot	(Atkinson, 1994)	`atkinsonstalactiteplot`

Installation and basic usage

`LinRegOutliers` can be downloaded and installed using the Julia package manager by typing

```
julia> using Pkg
julia> Pkg.add("LinRegOutliers")
```

in the Julia console. The regression methods follow a uniform call convention. For instance, a user can type

```
julia> setting = createRegressionSetting(@formula(calls ~ year), phones);
julia> smr98(setting)
Dict{String,Array{Int64,1}} with 1 entry:
  "outliers" => [15, 16, 17, 18, 19, 20, 21, 22, 23, 24]
```

or

```
julia> X = hcat(ones(24), phones[:, "year"]);
julia> y = phones[:, "calls"];
julia> smr98(X, y)
Dict{String,Array{Int64,1}} with 1 entry:
  "outliers" => [15, 16, 17, 18, 19, 20, 21, 22, 23, 24]
```

to apply *smr98* (Sebert et al., 1998) on the Telephone dataset (Rousseeuw, 1984), where X is the design matrix with ones in its first column. In this case, observations 15 to 24 are reported as outliers by the method. Some methods may also return additional information specific to the method which is passed back in a `Dict` object. For example, the *ccf* function returns a `Dict` object containing *betas*, *outliers*, *lambdas*, and *residuals*:

```
julia> ccf(X, y)
Dict{Any,Any} with 4 entries:
  "betas"     => [-63.4816, 1.30406]
  "outliers"  => [15, 16, 17, 18, 19, 20]
  "lambdas"   => [1.0, 1.0, 1.0, 1.0, 1.0, ...
  "residuals" => [-2.67878, -1.67473, -0.37067, -0.266613, …
```

Indices of outliers can be accessed using standard `Dict` operations like

```
julia> result = ccf(X, y)
julia> result["outliers"]
6-element Array{Int64,1}:
 15
 16
 17
 18
 19
 20
```

Acknowledgements

Guillermo Angeris is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656518.

References

Adnan, R., Setan, H., & Mohamad, M. N. (2000). *Identifying multiple outliers in linear regression: Robust fit and clustering approach*.

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, *89*(428), 1329–1339. https://doi.org/10.1080/01621459.1994.10476872

Barratt, S., Angeris, G., & Boyd, S. (2020). Minimizing a sum of clipped convex functions. *Optim Lett*, *14*, 2443–2459. https://doi.org/10.1007/s11590-020-01565-4

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571). John Wiley & Sons.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Billor, N., Chatterjee, S., & Hadi, A. S. (2006). A re-weighted least squares method for robust regression estimation. *American Journal of Mathematical and Management Sciences*, *26*(3-4), 229–252. https://doi.org/10.1080/01966324.2006.10737673

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, *34*(3), 279–298. https://doi.org/10.1016/S0167-9473(99)00101-2

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Chatterjee, S., & Mächler, M. (1997). Robust regression:a weighted least squares approach. *Communications in Statistics - Theory and Methods*, *26*(6), 1381–1394. https://doi.org/10.1080/03610929708831988

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18. https://doi.org/10.2307/1268249

Fischler, M. A., & Bolles, R. C. (1987). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In M. A. Fischler & O. Firschein (Eds.), *Readings in computer vision* (pp. 726–740). Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051581-6.50070-2

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(3), 761–771. https://doi.org/10.1111/j.2517-6161.1992.tb01449.x

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(2), 393–396. https://doi.org/10.1111/j.2517-6161.1994.tb01988.x

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, *88*(424), 1264–1272. https://doi.org/10.1080/01621459.1993.10476407

Hawkins, D. M., & Olive, D. (1999). Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis*, *32*(2), 119–134. https://doi.org/10.1016/S0167-9473(99)00029-8

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*, *32*(9), 929–946. https://doi.org/10.1080/02664760500163599

Kianifard, F., & Swallow, W. H. (1989). Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression. *Biometrics*, *45*(2), 571–585. https://doi.org/10.2307/2531498

Mahalanobis, P. C. (1930). *On tests and measures of group divergence. Part 1: Theoretical formulae* (Vol. 26, pp. 541–580). Journal & Proceedings Asiatic Society of Bengal (New Series).

Marchette, D. J., & Solka, J. L. (2003). Using data images for outlier detection. *Computational Statistics & Data Analysis*, *43*(4), 541–552. https://doi.org/10.1016/S0167-9473(02)00291-8

Nobakhti, A., Wang, H., & Tianyou Chai. (2009). Algorithm for very fast computation of least absolute value regression. *2009 American Control Conference*, 14–19. https://doi.org/10.1109/ACC.2009.5160229

Peña, D., & Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(3), 611–611. https://doi.org/10.1111/j.2517-6161.1995.tb02051.x

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880. https://doi.org/10.1080/01621459.1984.10477105

Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212–223. https://doi.org/10.1080/00401706.1999.10485670

Rousseeuw, P. J., & Van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In W. Gaul, O. Opitz, & M. Schader (Eds.), *Data analysis: Scientific modeling and practical application* (pp. 335–346). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-58250-9_27

Satman, M. H. (2015). Fast online detection of outliers using least-trimmed squares regression with non-dominated sorting based initial subsets. *International Journal of Advanced Statistics and Probability*, *3*(1), 53. https://doi.org/10.14419/ijasp.v3i1.4439

Satman, M. H. (2013). A new algorithm for detecting outliers in linear regression. *International Journal of Statistics and Probability*, *2*(3). https://doi.org/10.5539/ijsp.v2n3p101

Satman, M. H. (2012). A genetic algorithm based modification on the LTS algorithm for large data sets. *Communications in Statistics - Simulation and Computation*, *41*(5), 644–652. https://doi.org/10.1080/03610918.2011.598989

Sebert, D. M., Montgomery, D. C., & Rollier, D. A. (1998). A clustering algorithm for identifying multiple outliers in linear regression. *Computational Statistics & Data Analysis*, *27*(4), 461–484. https://doi.org/10.1016/s0167-9473(98)00021-8

Van Aelst, S., & Rousseeuw, P. (2009). Minimum volume ellipsoid. *WIREs Computational Statistics*, *1*(1), 71–82. https://doi.org/10.1002/wics.19