Author(s): Taskinen, Sara; Nordhausen, Klaus

Title: Iterative Weighted Least Squares

Year: 2022

Version: Accepted version (Final draft)

Please cite the original version:

Taskinen, S., & Nordhausen, K. (2022). Iterative Weighted Least Squares. In B. S. Daya Sagar, Q. Cheng, J. McKinley, & F. Agterberg (Eds.), Encyclopedia of Mathematical Geosciences. Springer. Encyclopedia of Earth Sciences Series. https://doi.org/10.1007/978-3-030-26050-7_169-1

Title: Iterative Weighted Least Squares

Name: S. Taskinen, K. Nordhausen

Affil./Addr.: Department of Mathematics and Statistics
University of Jyväskylä,
Jyväskylä, Finland
sara.l.taskinen@jyu.fi, klaus.k.nordhausen@jyu.fi

Iterative Weighted Least Squares

Definition

Iterative (re-)weighted least squares (IWLS) is a widely used algorithm for estimating regression coefficients. In the algorithm weighted least squares estimates are computed at each iteration step so that weights are updated at each iteration. The algorithm can be applied to various regression problems like generalized linear regression or robust regression. In this article we will focus however on its use in robust regression.

Introduction

Consider a data set consisting of n independent and identically distributed (iid) observations $(\boldsymbol{x}_i^\top, y_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ and y_i are the observed values of the predictor variables and the response variable, respectively. The data are assumed to follow the linear regression model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

where the p-vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ contains the unknown regression coefficients which are to be estimated based on the data. The errors ϵ_i are iid with mean zero and variance σ^2 and independent of $\boldsymbol{x}_i$.

The well-known least squares (LS) estimator for $\boldsymbol{\beta}$ is the $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals $\sum_i r_i(\boldsymbol{\beta})^2$, where $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}$, or equivalently, solves $\sum_i r_i(\boldsymbol{\beta})\, \boldsymbol{x}_i = \mathbf{0}$. To put this into a matrix form, let us collect the responses and predictors into a $n \times 1$ vector $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and a $n \times p$ matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, respectively, then the LS problem is given by

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2, \tag{1}$$

and the estimate can be simply computed as

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

assuming that $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists.

If the assumption of constant variance of the errors ϵ_i is violated, we can use the weighted least squares method to estimate the regression coefficients $\boldsymbol{\beta}$. The weighted least squares (WLS) estimate solves $\sum_i w_i\, r_i(\boldsymbol{\beta})\, \boldsymbol{x}_i = \mathbf{0}$, where $w_1, \ldots, w_n$ are some non-negative, fixed weights. If we let $\boldsymbol{W}$ be a $n \times n$ diagonal matrix with weights $w_1, \ldots, w_n$ on its diagonal, then the WLS estimate can be computed by applying ordinary least squares method to $\boldsymbol{W}^{1/2}\boldsymbol{y}$ and $\boldsymbol{W}^{1/2}\boldsymbol{X}$. Thus

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{y}, \tag{2}$$

If we assume that the errors ϵ_i are independent with mean zero and variance σ_i^2, then the WLS estimate with weights $w_i = 1/\sigma_i^2$ is optimal. If we further assume that the error distribution is Gaussian, then the WLS estimate is the maximum likelihood estimate. Notice that in practice the weights are often not known and need to be estimated. Some examples of weight selection are given, for example, in Montgomery et al (2012). Notice also that linear regression with non-constant error variance is only one application area of WLS.

Iterative weighted least squares

When the weight matrix $\boldsymbol{W}$ in (2) is not fixed, but may for example depend on the regression coefficients via residuals, we can apply the iterated weighted least squares (IWLS) algorithm for estimating the parameters. In such a case, the regression coefficients and weights are updated alternately as follows

1. Compute an initial regression estimate $\hat{\boldsymbol{\beta}}_0$.
2. For $k = 0, 1, \ldots$, compute the residuals $r_{i,k}(\hat{\boldsymbol{\beta}}_k) = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_k$ and weight matrix $\boldsymbol{W}_k = \mathrm{diag}(w_{1,k}, \ldots, w_{n,k})$, where $w_{i,k} = w(r_{i,k}(\hat{\boldsymbol{\beta}}_k))$ with some weight function w. Then update $\hat{\boldsymbol{\beta}}_{k+1}$ using WLS as in (2) with weight matrix $\boldsymbol{W}_k$, that is,
$$\hat{\boldsymbol{\beta}}_{k+1} = (\boldsymbol{X}^\top \boldsymbol{W}_k \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}_k \boldsymbol{y}.$$
3. stop when $\max_i |r_{i,k} - r_{i,k+1}| < \epsilon$, where ϵ may be fixed or related to the residual scale.

It is shown in Maronna et al (2018) that the algorithm converges if the weight function $w(x)$ is non-increasing for $x > 0$, otherwise starting values should be selected with care.

IWLS and robust regression

Let us next illustrate how IWLS is used in the context of robust regression, which is widely used in geosciences as atypical observations should not have an impact on the parameter estimation and often should be detected. For a recent comparison of non-robust regression with robust regression applied to geochemical data see for example van den Boogaart et al (2021). For a given estimate $\hat{\sigma}$ of the scale parameter σ, a robust M estimate of regression coefficient $\boldsymbol{\beta}$ can be obtained by minimizing

$$\sum_i \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right), \tag{3}$$

where ρ is a robust, symmetric ($\rho(-r) = \rho(r)$) loss function with a minimum at zero (Huber, 1981), or equivalently, by solving

$$\sum_i \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) \boldsymbol{x}_i = \mathbf{0}, \tag{4}$$

where $\psi = \rho'$. Here the scale estimate $\hat{\sigma}$ is needed in order to guarantee the scale equivariance of $\hat{\boldsymbol{\beta}}$. If the estimate is not known, it can be estimated simultaneously with $\hat{\boldsymbol{\beta}}$. For possible robust scale estimators see for example Maronna et al (2018).

The IWLS algorithm can be used for estimating robust M estimates as follows. Write $w(x) = \psi(x)/x$ and further $w_i = w(r_i(\boldsymbol{\beta})/\hat{\sigma})$. Then the M estimation equation in (4) reduces to

$$\sum_i w_i \, r_i(\boldsymbol{\beta}) \, \boldsymbol{x}_i = \mathbf{0},$$

and the robust M estimate of regression can be computed using the IWLS algorithm. For the comparison of robust M estimates computed using different algorithms, see Holland and Welsch (1977). For monotone $\psi(x)$ the IWLS algorithm converges to a unique solution given any starting value. Starting values however affect the number of iterations and should therefore be chosen carefully. Maronna et al (2018) advice to use the least absolute value (LAV) estimate as $\boldsymbol{\beta}_0$. If the robust scale is estimated simultaneously with the regression coefficients, it is updated at each iteration step. The median absolute deviation (MAD) estimate can then be used as a starting value for the scale. For a discussion on convergence in the case of simultaneous estimation of scale and regression coefficients, see Holland and Welsch (1977).

Some widely used robust loss functions include the Huber loss function

$$\rho_H = \begin{cases} \frac{1}{2}x^2, & |x| \le c \\ c\,(|x| - \frac{c}{2}), & |x| > c, \end{cases} \quad \text{yielding} \quad \psi_H(x) = \begin{cases} x, & |x| \le c \\ sign(x)\, c, & |x| > c, \end{cases}$$

and the Tukey biweight function

$$\rho_T = \begin{cases} 1 - \left[1 - \left(\frac{x}{c}\right)^2\right]^3, & |x| \le c \\ 1, & |x| > c, \end{cases} \quad \text{yielding} \quad \psi_T(x) = x\left[1 - \left(\frac{x}{c}\right)^2\right]^2 I(|x| \le c).$$

The tuning constants c provide a trade-off between robustness and efficiency at the normal model. Popular choices are $c_H = 1.345$ and $c_T = 4.685$ which yield an efficiency of 95% for the corresponding estimates. The two loss functions and their derivatives are shown in Figure 1.

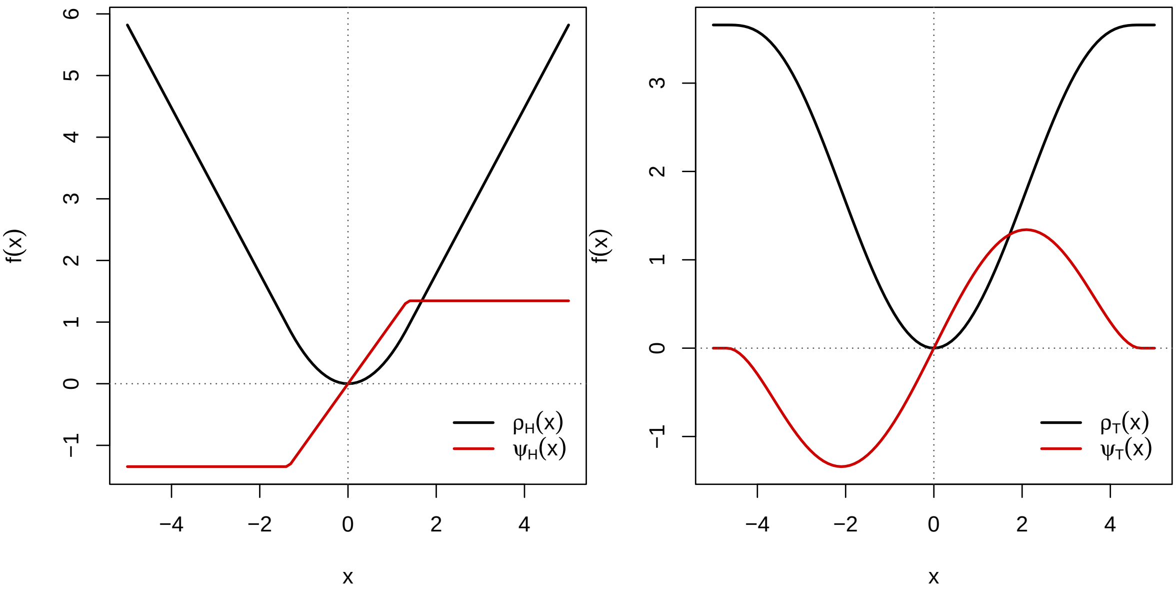

Fig. 1. Huber's and Tukey's biweight ρ and ψ functions with c chosen such that the efficiency at the normal model is 95%.

Notice that as the Tukey biweight function is not monotone, the IWLS algorithm may converge to multiple solutions. Good starting values are thus needed in order to ensure convergence to a good solution. The asymptotic normality of robust M regression estimates is discussed for example in Huber (1981).

Other usages of IWLS

Generalized linear models (GLM, McCullagh and Nelder, 1989), a generalization of the linear model described above, allows the response variable to come from the family of exponential distributions to which also the normal distribution belongs to. The regression coefficients in GLM are usually estimated via the maximum likelihood method which coincides with the LS method for normal responses. However, for other members from the exponential family there are no closed form expressions for the regression estimates. The standard way of obtaining the regression estimates is the Fisher scoring algorithm which can be expressed as an IWLS problem (McCullagh and Nelder, 1989). The Fisher scoring algorithm creates working responses at each iteration step. Therefore, when using IWLS for estimating coefficients of GLMs, not only are the weights in $\boldsymbol{W}$ updated at each iteration but also the (working) responses $\boldsymbol{y}$. How $\boldsymbol{W}_k$ and $\boldsymbol{y}_k$ are updated depends on the distribution of the response. For details, see for example McCullagh and Nelder (1989). The use of GLMs in soil science is for example discussued in Lane (2002).

The idea in robust M estimation described above is to down-weight large residuals whereas in the LS method, where the L_2 norm is used in the minimization problem (1), these get a large weight. Another way of approaching the estimation is to consider another norm, such as a general L_p norm, which then leads to

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_p^p,$$

which for example for $p = 1$ corresponds to the least absolute value (LAV) regression. It turns out that to solve the L_p norm minimization problem, also IWLS can be used with weights $\boldsymbol{W}_k = \operatorname{diag}(r_1(\boldsymbol{\beta}_k)^{2-p}, \ldots, r_n(\boldsymbol{\beta}_k)^{2-p})$. Uniqueness of the solution and behaviour of the algorithm depend on the choice of p. Also residuals which are too close to zero often need to be replaced by a threshold value. For more details, see for example Gentle (2007) and Burrus (2012). The motivation for using L_p norm in regression is that it may produce sparse solutions for $\boldsymbol{\beta}$ and thus can be used in image compression such as in hyperspectral imagery (see for example Zhao et al, 2020, for details).

Summary and Conclusion

The iterative weighted least squares algorithm is a simple and powerful algorithm which iteratively solves a least squares estimation problem. The algorithm is extensively employed in many areas of statistics such as robust regression, heteroscedastic regression, generalized linear models and L_p norm approximations.

Cross References

Ordinary Least Squares, Least Absolute Value, Least Mean Squares, Least Squares, Regression, Locally Weighted Scatterplot Smoother

References

van den Boogaart KG, Filzmoser P, Hron K, Templ M, Tolosana-Delgado R (2021) Classical and robust regression analysis with compositional data. Mathematical Geosciences 53(3):823–858, DOI 10.1007/s11004-020-09895-w

Burrus CS (2012) Iterative Reweighted Least Squares. OpenStax CNX

Gentle JE (2007) Matrix Algebra. Theory, Computations, and Applications in Statistics. Springer, New York

Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. Communications in Statistics - Theory and Methods 6(9):813–827

Huber PJ (1981) Robust statistics. Wiley, New York

Lane PW (2002) Generalized linear models in soil science. European Journal of Soil Science 53(2):241–251, DOI 10.1046/j.1365-2389.2002.00440.x

Maronna RA, Martin RD, Yohai VJ, Salibian-Barrera M (2018) Robust Statistics: Theory and Methods (with R), 2nd edn. Wiley, Hoboken

McCullagh P, Nelder J (1989) Generalized Linear Models, 2nd edn. Chapmann & Hall, Boca Raton

Montgomery DC, Peck EA, Vining GG (2012) Introduction to Linear Regression Analysis, 5th edn. Wiley & Sons, Hoboken

Zhao X, Li W, Zhang M, Tao R, Ma P (2020) Adaptive iterated shrinkage thresholding-based lp-norm sparse representation for hyperspectral imagery target detection. Remote Sensing 12(23):3991, DOI 10.3390/rs12233991