# Robust Linear Regression for Undrained Shear Strength Data

Jun Lin[1], Guojun Cai[1(✉)], Songyu Liu[1], and Anand J. Puppala[2]

[1] Institute of Geotechnical Engineering, Southeast University, Nanjing, Jiangsu, China
focuscai@163.com

[2] Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX, USA

**Abstract.** Outlier data has attracted considerable interesting geotechnical data. When doing classical linear least squares regression, if the regression data satisfied certain regression weights, the ordinary least squares regression is considered as the best method. However, the estimating and regression results may be inaccurate in case of these data not meeting given assumptions. Particularly in least squares regression analysis, there is some data (outliers) violating the assumption of normally distributed residuals. Under situation of regression data blending to outliers, robust regression is the best fit method. It can discriminate outliers and offer robust results when the regression data exists outliers. The purpose of this study is to make use of robust regression method to trend regression in geotechnical data analysis. Without defining absolute outliers from geotechnical testing data, outlier data of undrained shear strength is detected based on robust regression result.

**Keywords:** Undrained shear strength · Robust regression · Outlier data

## 1 Introduction

Geotechnical engineers face a number of uncertainties [1, 2]. Soil materials formed from geological weathering processes, and by physical means to deliver the soil to the current position [3]. In the forming process, the soil is influenced by various stress, pore fluid, and physical and chemical changes. Therefore, it is not surprising that there are always some outliers in geotechnical data. When dealing with geotechnical problems, empirical correlations between in situ or laboratory test results and geotechnical parameters are often used in geotechnical design. When establishing such empirical correlations, mostly adopted method is regression analysis, including linear or nonlinear regression [4].

Linear least squares regression (LLR) is a modeling approach by far the most widely used. When people say they use "regression", "linear regression" or "least squares" to adapt their data, they usually mean doing LLR. LLR is not only the most widely used method of modeling, but also have adapted to a variety of circumstances, beyond its immediate scope [5].

© Springer Nature Singapore Pte Ltd. 2018
L. Hu et al. (Eds.): GSIC 2018, *Proceedings of GeoShanghai 2018 International Conference: Multi-physics Processes in Soil Mechanics and Advances in Geotechnical Testing*, pp. 509–515, 2018.
https://doi.org/10.1007/978-981-13-0095-0_57

A mathematical method that finds the best-fit curve for a given set of points is to minimizing the sum of the squares of the distances of regression data deviating from the curve. The sum of squares of the offset distances is used instead of the absolute values of the offset distances because this allows the residuals to be treated as a continuously differentiable quantity. Whereas, because of the use of the square of the offset, peripheral points may have a disproportionate effect on fit. Whether the results are desirable or not, it depends on the issue of question [6].

The statistical observations of outliers are significantly different from the other sample values. Least-squares regression is obviously the best option if errors are normally distributed. Then, other means is eagerly required if these errors are not normally distributed. One particular distribution is the long tail error distribution of great concern. One solution is still to use the LLR method after removing the largest remaining value as outliers. However, this solution may be infeasible if several larger residual values exist by reason that the poor nature of the outlier tests. In addition, outlier testing is an acceptance/rejection process. The testing processes are neither smooth or statistically efficient. Robust regression (RR) is another option for least-squares regression in the case of the data contaminated with outliers. It can also be used to detect influential observations when the data is exposed to outliers [7].

It is difficult to define absolute outliers from geotechnical testing data, but it is possible to indicate the least predictable or relatively outlying data points using statistical tools. The objective of this paper is to demonstrate the advantages of RR analysis used in geotechnical data comparing with least square regression analysis. The procedure of RR is discussed shortly, based on LLR method. And then, Regression analysis is operated on undrained shear strength ($s_u$) data derived from CPTU test with both RR and LLR. Comparing regression result, the outlier of $s_u$ data can be detected based on RR method.

## 2 $s_u$ Data from CPTU

Unlike traditional sampling and laboratory tests, the piezocone penetration (CPTU) tests overcome the sampling disturbances with ease. In addition, comparing with conventional sampling and laboratory tests, CPTU tests can define an profile of $s_u$ along with depth with remarkable less time and effort. Besides, the profile of $s_u$ results is nearly continuous with depth, instead of at relatively few points of sampling and testing.

Various methods were proposed to determine $s_u$ results from CPTU data. Generally, these methods can be divided into theoretical and empirical relationships. The cone penetration into soils is a complex process. Because of the limitation of theoretical methods applied to simulating soil behavior during cone penetration process, empirical relationships are more favored to determine $s_u$ data in this study. These empirical relationships mainly include the direct or indirect correlations between cone penetration resistance and $s_u$. Another reason for adopting empirical determining relationships is to avoid too much on site and lab work.

A regular practice to determine to $s_u$ is to establish a relationship between $s_u$ and a net cone resistance. The net cone resistance is defined as $q_t - \sigma_{v0}$, where $\sigma_{v0}$ is the

in-situ total overburden stress. The equation links $s_u$ to the net cone resistance is given as:

$$s_u = \frac{q_t - \sigma_{v0}}{N_{kt}} \tag{1}$$

where $N_{kt}$ is an constant quantity. Numerous studies have been conducted to determine the proper values of $N_{kt}$. In this research, $N_{kt}$ is chosen to be 12.

## 3 Robust Linear Regression

### 3.1 Procedure of Robust Linear Regression

When the error distribution is not a normal distribution, the linear least squares estimation is not suitable, especially when the error has a heavy tail characteristic. The usual approach is to remove these relatively large weight data from the observed data in the least squares regression process. Another approach, so-called "robust regression," uses a more sophisticated approach that makes the method insensitive to outlier data. The most extensively used robust regression method is m-estimate. Such estimates can be viewed as a generalization of maximum likelihood estimates and is therefore called "m-estimate."

Considering the most generally linear regression model,

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \\ &= x_i' \beta + \varepsilon_i \end{aligned} \tag{2}$$

for the $i$ th of $n$ observations.

To estimate $b$ for $\beta$, the linear regression form is

$$\widehat{y}_i = \alpha + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + \varepsilon_i \tag{3}$$

and the residuals are given by

$$e_i = y_i - \hat{y}_i \tag{4}$$

In M-estimation method, the estimator $b$ is inferred by minimizing a specific objective function over all $b$,

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i' b) \tag{5}$$

where the specific function $\rho$ defines the weight of each residual in the specific function.

Let $\varphi = \rho'$ be the derivative of $\rho$. $\varphi$ is called the influence curve. Deriving the partial derivative of the objective function, combining with the coefficients $b$ and

setting the equation to zero and, then yields a coefficient estimation equation with $k+1$ rank:

$$\sum_{i=1}^{n} \varphi(y_i - x_i' b) x_i' = 0 \tag{6}$$

Define the weight function

$$\omega(e) = \frac{\varphi(e)}{e} \tag{7}$$

and let

$$\omega_i = \omega(e_i) \tag{8}$$

The equation of the estimated coefficient can be rewritten as

$$\sum_{i=1}^{n} \omega_i (y_i - x_i' b) x_i' = 0 \tag{9}$$

To solve these estimating equations is equivalent to a weighted least-squares regression, finding min $\sum \omega_i^2 e_i^2$.

However, the weights depend on the residuals, the residuals depend on the estimated coefficients, and the estimated coefficients depend on the weights. Therefore, the iterative weighting least squares (IRLS) is used to solve this problem: A solution (called iteratively reweighted least-squares), is therefore required:

Set least-squares estimates as initial estimates $b^{(0)}$.

In every iteration $t$, residuals $e_i^{(t-1)}$ are determined, and corresponding weights $\omega_i^{(t-1)}$ from the former iteration are also calculated.

To solve new weighted-least-squares estimates

$$b^{(t)} = \left[X' W^{(t-1)} X\right]^{-1} X' W^{(t-1)} y \tag{10}$$

where $X$ is the model matrix, with $x_i'$ as its $i$ th row, and $W^{(t-1)} = diag\left\{\omega_i^{(t-1)}\right\}$ is the current weight matrix.

Repeating step 2 and step 3 until the estimated coefficients tend to converge.

The asymptotic covariance matrix of $b$ is

$$v(b) = \frac{E(\varphi^2)}{[E(\varphi')^2]} (XX')^{-1} \tag{11}$$

Using $\sum[\varphi(e_i)]^2$ to estimate $E(\varphi^2)$, and $[\sum \varphi'(e_i)/n]^2$ to estimate $[E(\varphi')^2]$ produces the estimated asymptotic covariance matrix, $\hat{v}(b)$ (which is not reliable in small samples).

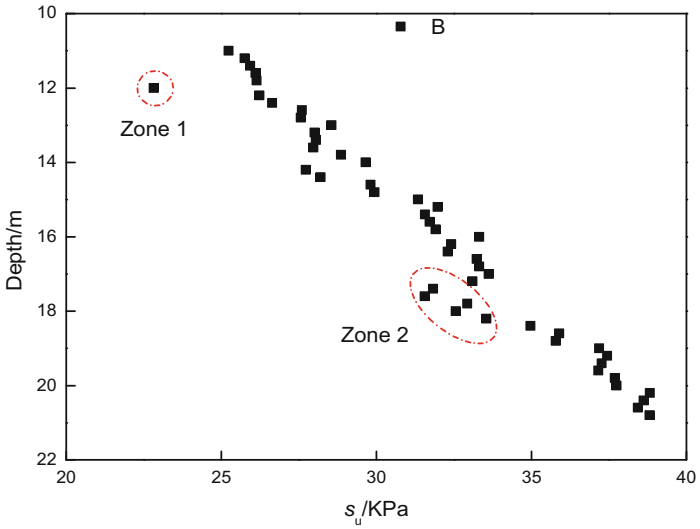


**Fig. 1.** $s_u$ data from CPTU test

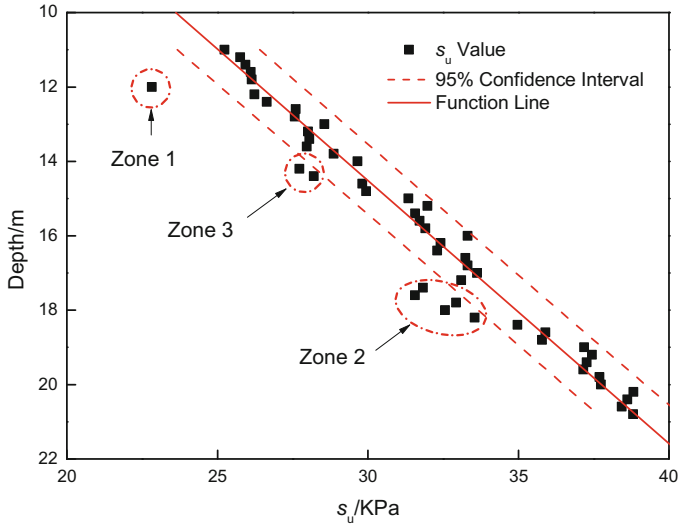


**Fig. 2.** RR result of $s_u$ data

### 3.2 Results of RR Regression and LLR Regression

A case study demonstrates the RR method on the $s_u$ data form CPTU test. The number of $s_u$ data used for regression is about 50. Visual inspection on the $s_u$ data, the outlier data is likely located in Zone 1 and Zone 2, shown in Fig. 1. To detect the outlier data and get a better function to describe the $s_u s$ data trend with depth, RR method and LLR method are used to regression analysis.

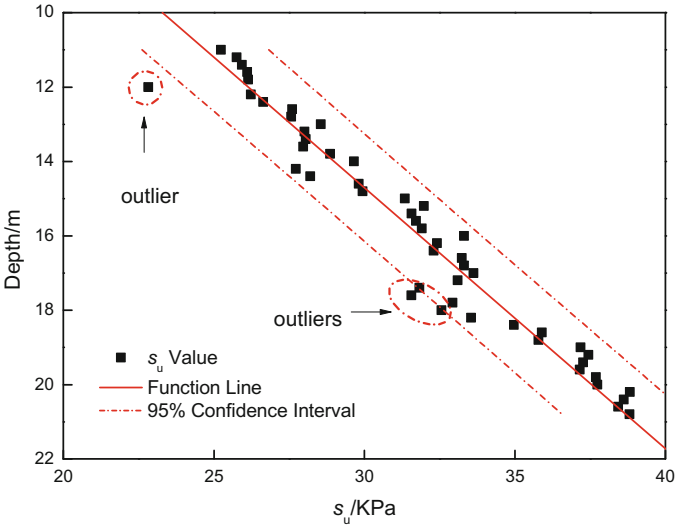


**Fig. 3.** LLR result of $s_u$ data

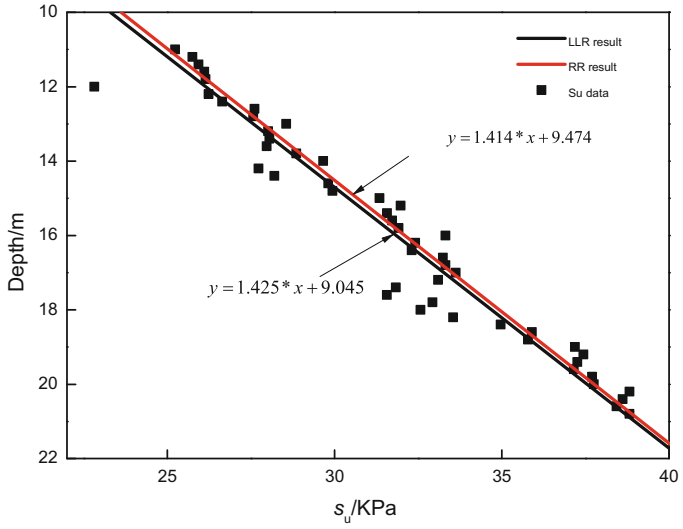


**Fig. 4.** Comparison of all results

The RR result is shown in Fig. 2 and LLR result in Fig. 3. Both results are provided with regression functions with 95% confidence interval. Take an inspection on the results in Figs. 2 and 3. It is obliviously that the RR regression function has a narrower 95% confidence interval (Fig. 2) than the LLR regression function (Fig. 3). And the RR regression function lies in the middle of $s_u$ data in the trend direction. The $s_u$ data scatters in both side of the RR regression function in equally chance. As a contrast, the LLR regression function has a wider 95% confidence interval. And the LLR crosses less $s_u$ data than RR regression in the trend direction. Most important is that the $s_u$ data scatters more in upper 95% confidence interval than in lower 95% confidence interval, as shown in Fig. 3. It means that the LLR is bias in such situation. As shown in Fig. 2, data in Zone 1 and Zone 2 and Zone 3 is more likely to be outlier data (Fig. 4).

## 4 Conclusion

This paper demonstrates the RR regression analysis and LLR analysis in the case that outlier data existed in the $s_u$ data. The regression analysis results show that RR method can deal with the outlier data in $s_u$ data very well. The RR regression function can give a more desirable result than the LLR function. Usually, the RR regression function has a narrower confidence interval than the LLR regression function. It is highly recommended that RR regression analysis should be adopted in the case that there is some outlier data existing in the geotechnical data.

## References

1. Zhang, L.M., Tang, W.H., Zhang, L.L., Zheng, J.G.: Reducing uncertainty of prediction from empirical correlations. J. Geotech. Geoenviron. Eng. **130**(5), 526–534 (2004)
2. Ching, J., Phoon, K.: Characterizing uncertain site-specific trend function by sparse Bayesian learning. J. Eng. Mech. **143**(7), 4017028 (2017)
3. Baecher, G.B., Christian, J.T.: Reliability and Statistics in Geotechnical Engineering. Wiley, New York (2003)
4. Gillins, D.T., Bartlett, S.F.: Multilinear regression equations for predicting lateral spread displacement from soil type and cone penetration test data. J. Geotech. Geoenviron. Eng. **141** (4), 04013047 (2015)
5. Yuen, K.V., Ortiz, G.A.: Outlier detection and robust regression for correlated data. Comput. Methods Appl. Mech. Eng. **313**, 632–646 (2016)
6. Gürünlü Alma, Ö.: Comparison of robust regression methods in linear regression. Int. J. Contemp. Math. Sci. **6**(9–12), 409–421 (2011)
7. Davies, P.L.: Aspects of robust linear regression. Ann. Stat. **21**(4), 1843–1899 (1993)