American Journal of Mathematics and Statistics 2017, 7(2): 71-77
DOI: 10.5923/j.ajms.20170702.03

Weighted Robust Lasso and Adaptive Elastic Net Method for Regularization and Variable Selection in Robust Regression with Optimal Scaling Transformations

Tarek M. Omara

Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Kafrelsheikh University, Kafrelsheikh, Egypt

Abstract In this paper, the weight least absolute deviation adaptive lasso optimal scaling method (WLAD-CATREG adaptive lasso) and weight least absolute deviation adaptive elastic net regression with optimal scaling method (WLAD-CATREG adoptive elastic net) will introduced, which is combined of weight least absolute deviation regression (WLAD-CATREG) and adaptive lasso (A-Lasso) or adaptive elastic net regression (A-Elastic net) with optimal scaling. Thus (WLAD-CATREG adoptive elastic net) method aim to automatically select variable, aspire to gropes effect and erase the bad effect of leverage points and outliers simultaneously, these aims cannot be achieved by (WLAD-CATREG), adaptive lasso regression (A-Lasso), weight robust adaptive lasso regression (WLAD-CATREG adoptive lasso), Weight least absolute deviation elastic net regression (WLAD-CATREG elastic net). Simulation study will be running to validated superiority of the (WLAD-CATREG adoptive Lasso) and (WLAD-CATREG adoptive elastic net).

Keywords A-Lasso, A-Elastic net, WLAD-CATREG, LAD-adoptive lasso, WLAD-CATREG adoptive lasso, WLAD-CATREG adoptive elastic net

1. Introduction

In regression models, if the variables are categorical variables, the relation between independent and dependent variables will be nonlinear. In this case, we used optimal scaling to transforming categorical variables to numeric variables, thus regression model becomes linear. This transform is done in simultaneous with the estimate parameters. The ordinary least squirt (OLS) is common method for estimating regression model but it sensitively to the outliers. In the case of regression model with optimal scaling transformations the effects of outliers can be as yet large (Peter (1993)), thus robust regression is suitable alternative. The robust regression include several methods one of them least absolute deviation (LAD) regression method which used to deal with outliers. This method developed to deal with leverage point (outlier in independent variables) by use weights which effect only on leverage point. In this side, weight least absolute deviation (WLAD) regression method proposed by (Giloni et al., 2006a, b, Olcay 2011). The regression model possibly is suffered from variable select problem, thus lasso regression method (Tibshirani (1996)) is appropriate because it does shrinkage parameter and variable selection simultaneously. The same tuning parameter was used in lasso regression for all coefficient, so it suffered an palpable bias and not have the oracle properties. (Fan and Li (2001)), therefor adaptive lasso regression (A-Lasso) was used to allow several tuning parameters for several coefficients ((Zou (2006)), Hansheng et al. (2006)). To avoid outlier, adaptive lasso (A-Lasso) objective function been modified in to least absolute deviation adaptive lasso regression (LAD-adoptive lasso) which has oracle properties when we appropriately chosen tuning parameter (Hansheng et al. (2006) and Xu and Yin (2010)) and to avoid leverage points, (LAD-adoptive lasso) combined of weight least absolute deviation regression (WLAD) and adaptive lasso regression (A-lasso) (Olcay, 2011).

The lasso regression has limitation when $p \gg n$, and if there is a group of highly correlated covariates, the lasso select one variable form the group, therefore (Zou and Hastie (2005)) introduce elastic net regression which basically combined penalty of ridge and lasso (L_1, L_2) and has not an oracle properties. The elastic net does shrinkage parameter, variable selection and aspire to group effect simultaneously. To improve the elastic net (Samiran, 2007) combined of the adaptive lasso and elastic net and get adaptive elastic net which have the oracle property when $p \gg n$.

In this paper, for categorical regression model, we introduce new estimators (WLAD- CATREG adoptive elastic net) which considered appropriate way to deal with

* Corresponding author:
tarek_em@yahoo.com (Tarek M. Omara)
Published online at http://journal.sapub.org/ajms

both leverage points and outliers, select variable and groping correlated variable simultaneously when p » n.

2. WLAD-CATREG Adoptive Lasso, WLAD-CATREG Adaptive Elastic Net

Consider the categorical regression model (CATREG)

$$\upsilon_r(y_i) = \sum_{j=1}^{k} \upsilon_j(x_{ij})\beta_j + \varepsilon_i \quad , i = 1,2,\dots,n \qquad (1)$$

Where y is the discretized response variable, x_i is the discretized predictor variables, υ_r is the function of transformation response variable, υ_j is the function of transform predictor variables, β_j are the regression coefficients and ε is independent random error. The form of the last transformation based on optical scaling, so in the case of numerical scaling level the result of CATREG is the same of standardized multiple linear regression (Anita and van (2007)).

The (OLS) is the more commonly used for estimating the previous model. But this method is poorly when the model have outliers, thus many robust methods introduced. One important of the robust method called L_pCATREG regression methods. Since y_i and x_{ij} are standardized variables and if we fixed $\upsilon_r(y_i)$, $\upsilon_j(x_{ij})$ and β_l for all predictors l≠j then L_p is written as

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n}\left|\upsilon_r(y_i) - \sum_{j=1}^{k}\upsilon_j(x_{ij})\beta_j\right|^{p}\right], \quad p \geq 1 \qquad (2)$$

In the special case, (p=1), we get LAD CATREG or least absolute deviations regression with optimal scaling method

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n}\left|\upsilon_r(y_i) - \sum_{j=1}^{k}\upsilon_j(x_{ij})\beta_j\right|\right] \qquad (3)$$

The good leverage points, observation in the space of independent predictors, not affects significantly the LAD method, but bad leverage points, observation in the space of independent predictors but existing far from fit line, affects of it (Rousseeuw and Leroy (1987)). To deal with bad leverage points (Eills and Morgenthaler (1992)) suggested weight least absolute deviation (WLAD) regression method. An extinction of (WLAD) regression method, we get (WLAD-CATREG) regression method

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n} w_i\left|\upsilon_r(y_i) - \sum_{j=1}^{k}\upsilon_j(x_{ij})\beta_j\right|\right] \qquad (4)$$

Where $0 < w_i \leq 1$, w_i i=1,2,… ,n is the weights which will be chosen to breakdown leverage points. Chatterjee and Hadi, (1988) suggested weight depended of clean subset where

$$w_i = \sqrt{\frac{\min_j(h_j)}{h_I}} \quad ,i=1,2,...,n,$$

$$h_i = \upsilon(x^i)\left(\upsilon(X_C)^T\upsilon(X_C)\right)^{-1}\upsilon(x^i)^T$$

$\upsilon(X_C)$ is clean subset, $\upsilon(x^i)$ is the set of observation relative to clean subset. The discreet choice of weight lead to (WALD) estimator with fast computationally and high breakdown point. (Rousseeuw and Hubert (1997), (Olcay, 2011).) Used the robust distances (RD) to compute weights w_i which defined as

$$w_i = min\left\{1, \frac{k}{[RD(x_i)]^2}\right\} = min\left\{1, \frac{k}{\left[\sqrt{(x_i-\hat{\mu})\hat{\Sigma}^{-1}(x_i-\hat{\mu})^T}\right]^2}\right\} \qquad (5)$$

Where $\hat{\mu}, \hat{\Sigma}$ are location and scatter estimators. Since robust distances (RD) identify leverage points increasing, (RD) lead to decrease weights w_i and thus leverage points corresponds to similar weights then it will be down weighted.

(Tibshirani (1996)) introduce lasso regression to combines estimate and variable selection. The lasso regression depend on minimize least squirt regression with the L_1 norm condition. An extinction lasso regression, (Anita and van (2007)) introduce lasso penalties with (CATREG)

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n}\left[\upsilon_r(y_i) - \sum_{j=1}^{k}\upsilon_j(x_{ij})\beta_j\right]^2 + \lambda_1\sum_{j=1}^{k}|\beta_j|\right] \qquad (6)$$

The lasso regression not have the oracle property, so ((Zou (2006)) introduce adaptive lasso (A-Lasso) which have oracle property when p » n. The adaptive lasso (A-Lasso) with (CATREG) defined as

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n}\left[\upsilon_r(y_i) - \sum_{j=1}^{k}\upsilon_j(x_{ij})\beta_j\right]^2 + \lambda_1\sum_{j=1}^{k} w_{s_j}|\beta_j|\right] \qquad (7)$$

Where w_{s_j}, w_{s_l} is two weight vector, $w_{s_j}, w_{s_l} > 0$. In order to reach oracle property, ((Zou (2006), Olcay (2011)) define the weight vector as $\hat{w}_{s_j} = |\hat{\beta}_j|^{-\gamma}$, j=1,2,…,k, Where γ is a positive constant and $\hat{\beta}_j$ is an elastic net estimater of β_j. (Zou and Zhang (2009)) used the other formula for weight vector as $\hat{w}_{s_j} = |\hat{\beta}_j + 1/n|^{-\gamma}$, j=1,2,…,k and (Wang et al (2007)) choose w_{s_j} as

$$\hat{w}_{sj} = \left[\log(n)/n|\hat{\beta}_j|\right]^{-1} j = 1,2, \dots,k \qquad (8)$$

which satisfies $\sqrt{n}\,\hat{w}_{sj} \to 0$ for $j \leq k_0$ and $\sqrt{n}\,\hat{w}_{sj} \to \infty$ for $j > k_0$ to avoid dividing zeros.

Since least squirt errors sensitive to outliers, (Hansheng and Chenlei. (2006)), introduced least absolute deviation

adaptive lasso regression (LAD-adoptive lasso). In this way, (Olcay (2011)) introduce weight least absolute deviation regression (WLAD) and adaptive lasso regression (A-lasso) to avoid leverage point add to the features of (LAD-adoptive lasso) which produce robust parameter with oracle property and select variables. The (WLAD-adoptive lasso) with (CATREG) (New) get as

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n} w_i \left|\upsilon_r(y_i) - \sum_{j=1}^{k} \upsilon_j(x_{ij})\beta_j\right| + n\sum_{j=1}^{k} \lambda_{1_j}|\beta_j|\right] \quad (9)$$

On the other hand, (Zou and Hastie (2005)) introduce elastic net regression which add the group effect to the features of lasso regression. The elastic net regression depend on minimize least squirt regression with the L_1, L_2 norm conditions. (Anita and van (2007)) introduce naive elastic net penalties with (CATREG) (CATREG naïve elastic net).

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n}\left[\upsilon_r(y_i) - \sum_{j=1}^{k} \upsilon_j(x_{ij})\beta_j\right]^2 + \lambda_1\sum_{j=1}^{k}|\beta_j| + \lambda_2\sum_{j=1}^{k}\beta_j{}^2\right] \quad (10)$$

(Samiran, 2007) introduce (Adaptive- elastic net) which have oracle property. We get (Adaptive- naive elastic net) with (CATREG) as

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n}\left[\upsilon_r(y_i) - \sum_{j=1}^{k} \upsilon_j(x_{ij})\beta_j\right]^2 + \lambda_1\sum_{j=1}^{k} w_{s_j}|\beta_j| + \lambda_2\sum_{j=1}^{k}\beta_j{}^2\right] \quad (11)$$

Hong and Zhang (2010) developed the Adaptive naive elastic net to weight L_1 and L_2 penalty, thus we defined the adaptive elastic net with (CATREG) (New), in this case as

$$\beta = \operatorname{argmin}\frac{1}{2}\left[\sum_{i=1}^{n}\left[\upsilon_r(y_i) - \sum_{j=1}^{k} \upsilon_j(x_{ij})\beta_j\right]^2 + n\lambda_1\sum_{j=1}^{k} w_{s_j}|\beta_j| + \frac{n}{2}\lambda_2\sum_{j=1}^{k} w_{s_j}\beta_j{}^2\right] \quad (12)$$

The (LAD-CATREG adaptive elastic net (New)) which is scaled version of the (WLAD-CATREG naive adaptive elastic net (New)) defined as

$$\beta = \operatorname{argmin}\frac{1}{2}\left[\sum_{i=1}^{n}\left|\upsilon_r(y_i) - \sum_{j=1}^{k} \upsilon_j(x_{ij})\beta_j\right| + n\lambda_1\sum_{j=1}^{k} w_{s_j}|\beta_j| + \frac{n}{2}\lambda_2\sum_{j=1}^{k} w_{s_j}\beta_j{}^2\right] \quad (13)$$

When p » n, to avoid leverage point and outliers in the elastic net regression, we defined (WLAD-CATREG naive adaptive elastic net (New)) as

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n} w_i \left|\upsilon_r(y_i) - \sum_{j=1}^{k} \upsilon_j(x_{ij})\beta_j\right| + n\sum_{j=1}^{k} \lambda_{1_j}\left|w_{s_j}\beta_j\right| + \lambda_2\sum_{j=1}^{k} w_{s_j}\beta_j{}^2\right] \quad (14)$$

And we defined (WLAD-CATREG naive adaptive elastic net (New)) as

$$\beta_{\text{WLAD}-\text{CATREG naive adaptive elastic net}} = \left(1 + \boldsymbol{\lambda_2}^{(\mathbf{n})}\right)\beta_{\text{WLAD}-\text{CATREG adaptive elastic net}} \quad (15)$$

3. Algorithm

To simplify (13), let

$\upsilon_r(y_i)^* = w_i\upsilon_r(y_i)$, $\upsilon_r(\ddot{y}_i) = \begin{pmatrix}\upsilon_r(y_i)^*\\ 0\end{pmatrix}$,

$\upsilon_j(\tilde{x}_{ij}) = w_i\upsilon_j(x_{ij})$, $\ddot{\beta}_j = \sqrt{(1+\lambda_1)}\beta_j$,

and define

$\left(\upsilon_r(\ddot{y}_i), \upsilon_j(\ddot{x}_{ij})\right)$, i=1,2,…,n,n+1,…,n+k, j=1,2,…,k.

where

$\left(\upsilon_r(\ddot{y}_i), \upsilon_j(\ddot{x}_{ij})\right) = \left(\upsilon_r(y_i)^*, \upsilon_j(\tilde{x}_{ij})\right)$

$= \left(w_i\upsilon_r(y_I), w_i\upsilon_j(x_{ij})\right)$,

i=1,2,…,n, j=1,2,…,k,

$\left(\upsilon_r(\ddot{y}_{n+j}), \upsilon_j(\ddot{x}_{(n+j)\,j})\right) = (0, n\lambda_{1j}e_j)$

e_j is a p-dimension vector with jth term equals one and all others equal to zero. We can rewrite (13) as a form of ridge as

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n+k}\ddot{W}_i\left|\begin{pmatrix}\upsilon_r(y_i)^*\\ 0\end{pmatrix} - \sum_{j=1}^{k}\frac{1}{\sqrt{(1+\lambda_1)}}\begin{pmatrix}\upsilon_j(\tilde{x}_{ij})\\ w_{sj}\lambda_1\end{pmatrix}\sqrt{(1+\lambda_1)}\beta_j\right| + \frac{\lambda_2}{(1+\lambda_1)}\sum_{j=1}^{k}(1+\lambda_1)|w_{sj}\beta_j|^2\right]$$

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n+k}\ddot{W}_i\left|\upsilon_r(\ddot{y}_i) - \sum_{j=1}^{k}\upsilon_j(\ddot{x}_{ij})\ddot{\beta}_j\right| + \gamma_{ss}\sum_{j=1}^{k}|w_{sj}\ddot{\beta}_j|^2\right]$$

Where $\gamma_{ss} = \lambda_2/(1+\lambda_1)$.

Define $\ddot{X}_{ij_w} = \frac{\ddot{X}_{ij}}{w_{sj}}$, $\ddot{\beta}_{j_w} = w_s\ddot{\beta}_j$

Then

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n+k} \ddot{W}_i \left| \upsilon_r(\ddot{y}_i) - \sum_{j=1}^{k} \upsilon_j\left(\ddot{X}_{ij_w}\right)\ddot{\beta}_{j_w}\right| + \gamma_{ss}\sum_{j=1}^{k}\left|\ddot{\beta}_{j_w}\right|^2\right] \quad (16)$$

Where $\ddot{W}_i = \frac{1}{\left|\upsilon_r(\ddot{y}_i) - \sum_{j=1}^{k}\upsilon_j\left(\ddot{X}_{ij_w}\right)\ddot{\beta}_{j_w}\right|}$.

Before computing the WLAD-CATREG adaptive elastic net we must choose the weights w_i, w_{s_j} by use (and the tuning parameters λ_1, λ_2 can be chosen Cross-validation (CV) on a two-dimensional but it would be computationally prohibitive (See Li and Jia (2010)).

So that we fixed λ_2 and we used five-fold Absolute Cross-validation (ACV) to select tuning parameter λ_1 which avoided leverage point. The Absolute Cross-validation (ACV) defined as

$$\mathrm{ACV}(\lambda_2) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{1 - H(\lambda_2)_{ii}}\right|$$

Where H_{ii} is hat matrix, $H(\lambda_2) = X(X^TX + \lambda_2 I)^T X^T$. We chosen weights w_i by the robust distances (RD) which defined as (5).

We computation the WLAD-CATREG adaptive elastic net by development the algorithm:

Agorithm 1

1. Input: Design Matrix X.
2. Find: The robust estimates $\hat{\mu}, \hat{\Sigma}$.
3. Calculate the initial regression estimator $\hat{\beta}_j$ and the weight coefficient $\hat{w}_s, \hat{w}$.
4. Define $r(y_i)^* = \hat{w}_{ir}(y_i)$, $j(\tilde{x}_{ij}) = \hat{w}_{ij}(x_{ij})$.
5. Solve the original WLAD-CATREG adaptive lasso

$$\beta = \operatorname{argmin}\left[\sum_{i=1}^{n} w_i \left| r(y_i) - \sum_{j=1}^{k} j(x_{ij})\beta_j\right| + n\sum_{j=1}^{k}\lambda_{1j}|\beta_j|\right]$$

and the WLAD-CATREG naive adaptive elastic net

$$\hat{\beta} = \operatorname{argmin}\left[\sum_{i=1}^{n+k} \hat{w}_i \left| r(\ddot{y}_1) - \sum_{j=1}^{k} j\left(\ddot{X}_{ij_w}\right)\ddot{\beta}_{j_w}\right| + \gamma_{ss}\sum_{j=1}^{k}\left|\ddot{\beta}_{j_w}\right|^2\right]$$

6. Calculate WLAD-CATREG adaptive elastic net

$$\hat{\boldsymbol{\beta}}_{\mathrm{WLAD-CATREG\ adaptive\ elastic\ net}} = \left(\mathbf{1} + \boldsymbol{\lambda_2}^{(n)}\right)\hat{\beta}$$

Output:

$\hat{\boldsymbol{\beta}}_{\mathrm{WLAD-CATREG\ adaptive\ lasso}}$

$\hat{\boldsymbol{\beta}}_{\mathrm{WLAD-CATREG\ naive\ adaptive\ elastic\ net}}$,

$\hat{\boldsymbol{\beta}}_{\mathrm{WLAD-CATREG\ adaptive\ elastic\ net}}$

for j=1,2,…,k

4. The Properties of Estimators

In this section, we discuss the asymptotic properties of the WLAD-CATREG adaptive adaptive elastic net At the first, we rooting some convenience and definitions. We decompose the β vector as

$$\beta = (\beta_a^T, \beta_b^T)^T = [(\beta_1, \beta_2, \dots, \beta_{K0})^T, (\beta_{K0+1}, \dots, \beta_K)^T]^T,$$

decompose the predictor variables x_I as

$$\left(\upsilon_j(x_{ia}), \upsilon_j(x_{ib})\right) = \left[\left(\upsilon_j(x_{i1}), \upsilon_j(x_{i2}),\right), \dots, \upsilon_j(x_{ik0})\right)^T, \left(\upsilon_j\left(x_{i(k0+1)}\right), \dots, \upsilon_j(x_{ik})\right)^T\right]$$

Defined $a_n = max\left[\lambda_j : 1 \le j \le k0\right]$ and $b_n = min\left[\lambda_j : k0 < j \le k\right]$ where λ_j is a function of n. (Wang and Leng (2007)).

Let $\hat{\beta}_w = \left(\hat{\beta}_{ewa}^T, \hat{\beta}_{ewb}^T\right)^T$ be identical WLAD-CATREG adoptive net elastic estimator. Consider the linear regression model (1) with independent and assume the following conditions:

A1: The identically error with median zero and cumulative distribution function F which is positive and continues,

A2: The covariance of predictor variables x_i exists and positive definite

A3: W is n×n diagonal matrix with known positive value (w_i ,i=1,2,...,n), max w_i =O(1) and max w_i^{-1} = O(1)

A4: $\lim_{n\to\infty}\frac{X^TWX}{n} = \left(Q_n + \frac{\gamma_s^{(n)}}{n}I\right) \to Q$

where Q is a positive definite.

A5: $\sqrt{n}\lambda_1 = o(1)$, $\sqrt{n}\lambda_2 = o(1)$

A6: $\lim_{n\to\infty}\frac{\gamma_{ss}}{\sqrt{n}}\sqrt{\sum_{j\in\mathcal{A}}\left|\ddot{\beta}_{j_w}\right|^2} = 0$

The assumes (A1, A3, A4) are the same assumes for Olcay (2011), the assume (A4) is the (A6) assume for (Zou and Zhang (2009)) and the assumes A1, A2 the same assume for (Pollard, 1991).

Lemma (1): For the model (1), if it satisfies Assumptions A1: A6, then LAD-CATREG adaptive elastic net $\hat{\beta}^T = \left(\hat{\beta}_a^T, \hat{\beta}_b^T\right)^T$ (14) must satisfies the following:

$$P\left(\hat{\beta}_b = 0\right) \to 1$$

$$\sqrt{n}\left(\hat{\beta}_a^T - \beta_a\right) \to N\left(0, \Sigma_{w0}^{-1}/4f^2(0)\right)$$

Proof lemma (1):

Let $\ddot{\beta}_{j_w} = \tilde{\beta}_{j_w} + \frac{\tau_j}{\sqrt{n}}$

Then, we can rewrite (14) as

$$\Psi(\tau) = \sum_{i=1}^{n+k}\left|\ddot{y}_i - \sum_{j=1}^{k}\ddot{X}_{ij_w}\left(\tilde{\beta}_{j_w} + \frac{\tau_j}{\sqrt{n}}\right)\right| + \gamma_{ss}\sum_{j=1}^{k}\left|\left(\tilde{\beta}_{j_w} + \frac{\tau_j}{\sqrt{n}}\right)\right|^2$$

$$\tau_j = \sqrt{n}\left(\tilde{\beta}_{j_w} - \ddot{\beta}_{j_w}\right)$$

$$\Psi(0)=\sum_{i=1}^{n+k}\left|\ddot{y}_i-\sum_{j=1}^{k}\ddot{X}_{ij_w}\tilde{\beta}_{j_w}\right|+\gamma_{ss}\sum_{j=1}^{k}\left|\tilde{\beta}_{j_w}\right|^2$$

Let $D_n(\tau)=\Psi(\mu)-\Psi(0)=$

$$\sum_{i=1}^{n+k}\left|\ddot{y}_i-\sum_{j=1}^{k}\ddot{X}_{ij_w}\left(\tilde{\beta}_{j_w}+\frac{\tau_j}{\sqrt{n}}\right)\right|-\sum_{i=1}^{n+k}\left|\ddot{y}_i-\sum_{j=1}^{k}\ddot{X}_{ij_w}\tilde{\beta}_{j_w}\right|+\gamma_{ss}\sum_{j=1}^{k}\left|\left(\tilde{\beta}_{j_w}+\frac{\tau_j}{\sqrt{n}}\right)\right|^2-\gamma_{ss}\sum_{j=1}^{k}\left|\tilde{\beta}_{j_w}\right|^2$$

$$=\sum_{i=1}^{n+k}\left\{\left|\ddot{y}_i-\sum_{j=1}^{k}\ddot{X}_{ij_w}\left(\tilde{\beta}_{j_w}+\frac{\tau_j}{\sqrt{n}}\right)\right|-\left|\ddot{y}_i-\sum_{j=1}^{k}\ddot{X}_{ij_w}\tilde{\beta}_{j_w}\right|\right\}+\gamma_{ss}\sum_{j=1}^{k}\left\{\left|\left(\tilde{\beta}_{j_w}+\frac{\tau_j}{\sqrt{n}}\right)\right|^2-\left|\tilde{\beta}_{j_w}\right|^2\right\}$$

$$\geq\sum_{i=1}^{n+k_0}\left\{\left|\ddot{y}_i-\sum_{j=1}^{k_0}\ddot{X}_{ij_w}\left(\tilde{\beta}_{j_w}+\frac{\tau_j}{\sqrt{n}}\right)\right|-\left|\ddot{y}_i-\sum_{j=1}^{k_0}\ddot{X}_{ij_w}\tilde{\beta}_{j_w}\right|\right\}+na_n\sum_{j=1}^{k_0}\left\{\left|\tau_j\right|^2\right\}$$

$$\geq Z_n^{(1)}(\tau)+Z_n^{(2)}(\tau) \qquad (A1)$$

Knight (1998) holds that for $x\neq 0$

$$|x-y|-|x|=-y[I(x>0)-I(x<0)]+2\int_0^Y[I(x\leq s)-I(x\leq 0)]ds$$

Using this equation, the first item at (A1) $Z_n^{(1)}(\tau)$ can be expressed as

$$-n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_j\ddot{X}_{ij_w}[I(\ddot{\varepsilon}_i>0)-I(\ddot{\varepsilon}_{i_i}<0)]+2\sum_{i=1}^{n}\sum_{j=1}^{k}\int_0^{n^{-\frac{1}{2}}\mu_j\ddot{X}_{ij_w}}[I(x\leq s)-I(x\leq 0)]ds \qquad (A2)$$

Using the central limit theorem, the first item converges in distribution to $\mu^T H$, where $\mu^T=(\mu_1,\mu_2,\ldots,\mu_k)^T$ and H p-dimensional normal random vector with mean 0 and variance matrix

$\Sigma_w=\mathrm{cov}(\ddot{X}_w)$. Since $\lambda_1=o(n^{-1/2})$, $\max wi=O(1)$ and $\max w_i^{-1}=O(1)$ then $\Sigma_w=X^TW^2X$.

Follow from the proofs of lemma1 (Wang et al. (2007)), the second item converges to

$$f(0)\tau^T\Sigma_w\tau=f(0)\tau^TX^TW^2X\tau$$

$$Z_n^{(1)}(\tau)\underset{d}{\rightarrow}\tau^TH+f(0)\tau^TX^TW^2X\tau$$

Since $\sqrt{n}a_n\rightarrow 0$, the second item at (A1) $Z_n^{(2)}(\tau)$ converges to 0 in probability.

Then

$$D_n(\tau)\underset{d}{\rightarrow}\mu^TH_0+f(0)\tau^TX^TW^2X\tau$$

Where H_0 p-dimensional normal random vector with mean 0 and variance matrix Σ_{w0}.

Follow from the proofs of lemma1 (Wang et al. (2007)).

5. The Simulation Study

In this section, the performance of the WLAD-CATREG adaptive elastic net estimates are examined via the simulation. We simulate data sets from the true model $\upsilon_r(y_i)=\upsilon(x_i)^T\beta_j+\sigma\varepsilon_I$ where $\upsilon(x_i)=[\upsilon(x_{i1}),\upsilon(x_{i2}),\ldots\upsilon(x_{ik})]^T$. To ensure the situation of grouping variable, the simulated date consist of a training set, an independent validation set and an independent test set. In this simulate, we use the 50 simulated data sets each consisting of (30 training set/30 independent validation set /100 independent test set) observations and have 30 variables, 10 categorical variables each of them containing nine category and 20 numerical variables. The numerical variables generate $X_{2i}\sim N_k(2,I),i=1,2,\ldots,m$ recalling to the model $\upsilon_r(y_{2i})=\sum_{j=1}^{k}\upsilon_j(x_{2ij})\beta_{2j}$ where $\beta_{1j}\neq\beta_{2j}$ and the categorical variables generate by Bernoulli distribution with nine category. The correlation between pairs of variables $r=\mathrm{corr}(X_I,X_j),i\neq j$ was taken (0.20, 0.60). We have used WLAD-CATREG-l algorithm to compute the LAD-adaptive lasso estimator and WLAD-CATREG adaptive lasso estimator and WLAD-CATREG-EN algorithm to compute the LAD-CATREG adaptive elastic net estimator and WLAD-CATREG adaptive elastic net estimator. To contaminate the data, we generated the contamination rate (ë= 20%, 40%). We will choose the place of outliers from observations on are works well. And the elastic net -CATREG methods select more variables than the lasso methods. In adding, the WLAD naïve adoptive elastic net-CATREG make improvements to a number of selected variables. categorical variable randomly and replacing the category on the variables by extreme values (such as 1 or 9) and will generate the outliers on numerical variable by $\upsilon(x_{1i})\sim N_k(0,0.5^{|i-j|}),i=1,2,\ldots,n-m$ where $i\neq j$ recalling to the model $\upsilon_r(y_{1i})=\sum_{j=1}^{k}\upsilon_j(x_{1ij})\beta_{1j}+\sigma\varepsilon_I$.

To ensure the achievement the normality of error with the possibility of the existence of outlier of y, we generate ε from t distribution with two degrees of freedoms ε~ t(2) and set $\sigma=3$. We let

$$\beta_l=(\underbrace{1,1,1,1,1,1,0,0,0,0,0,0,2,2,2,2,2,2,3,3}_{,3,3,3,3,1,1,1,1,1,1)})$$

and for the weight vector, we assume $\gamma=5$ Since elastic parameters have two tuning parameters, we used 2-dimensional cross-validation. For this method, we use $\lambda_2=0.2$.

Table (1). ACV(λ_2) and number of nun- zero coefficients for the simulation study

	ACV($\lambda_2 = 0.2$)				**Number of nun- zero coefficients**			
***Estimators**	$r = 0.2$	$r = 0.6$	$r = 0.2$	$r = 0.6$				
	ë=20%	*ë=40%*	*ë=20%*	*ë=40%*	*ë=20%*	*ë=40%*	*ë=20%*	*ë=40%*
(1)	62.6	60.3	63.6	62.3	---	---	---	---
(2)	61.3	60.2	61.5	59.8	---	---	---	---
(3)	63.5	61.4	63.5	60.8	12	13	12	11
(4)	60.6	59.2	59.3	603	12	11	11	11
(5)	55.2	54.1	53.1	52.5	14	13	12	12
(6)	51.0	52.3	52.1	53.1	16	15	14	13
(7)	50.2	48.3	49.2	48.2	15	15	14	14
(8)	40.5	41.3	38.2	37.2	17	16	14	14
(9)	41.2	39.2	37.2	36.1	21	20	18	16

*(1) LAD- CATREG, *(2) WLAD-CATREG, *(3) Lasso-CATREG, *(4) adaptive lasso-CATREG,
*(5) WLAD adoptive lasso- CATREG, *(6) naïve elastic net- CATREG, *(7) Adaptive naive elastic net-CATREG,
*(8) weighted naïve elastic net-CATREG, *(9) WLAD naïve adoptive elastic net-CATREG.

In Table (1), the WLAD-CATREG method has a poor performance especially at high level of correlation and lasso-CATREG has working poor at high level of outliers. Parallel, the weighted naïve elastic net-CATREG and WLAD naïve adoptive elastic net-CATREG method have a best performance. In all cases, elastic net methods are more improved than lasso methods and the weight was given to improve the estimate methods. When the level of correlation and the level of outliers are increase, the elastic estimators.

6. Conclusions

We presented a new estimator for categorical regression model (CATREG) which takes into account effect the outliers, leverage point and variable select problem. The new criterion, Absolute Cross-validation (ACV) which use Cross-validation at robust form, was used for trade-off between estimators. The result for simulation study showed that, the weighted naïve elastic net-CATREG and the WLAD naïve adoptive elastic net-CATREG make improvements for the estimate.

REFERENCES

[1] Peter, V. (1993), "M-estimators in multiple regression with optimal scaling", Netherlands organization for scientific research, RR-92-10.

[2] Giloni, A., Simonoff, J., Sengupta, B., (2006a), "Robust weighted LAD regression", Computional Statistics & Data Analysis, Vol.50.

[3] Giloni, A., Sengupta, B., Simonoff, J., (2006b), "A mathematical programming approach for improving the robustness of least sum of ansolute deviations regression", Wiley InferScience doi, Vol.10.

[4] Olcay A., (2011), "Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression", Computational Statistics & Data Analysis, Vo.56.

[5] Tibshirani, R. (1996), "Regression shrinkage and selection Via the lasso", Journal of Royal Statistical Society B, Vo.58.

[6] Fan, J., and Li, R., (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American Statistical Association, Vo. 96.

[7] Zou, H., (2006), "The adaptive Lasso and its oracle properties", Journal of the American Statistical Association", Vo. 101.

[8] Hansheng W., Chenlei L., (2006), "A note on adaptive group lasso", Computational Statistics and Data Analysis, Vo.52.

[9] Xu J., Yi Z., (2010), "Simultaneous estimation and variable selection in median regression using Lasso-type penalty", Annals of the institute of Statistical Mathematics, Vo.62.

[10] Zou, H. and Hastie, T., (2005), "Regularization and variable selection via the elastic net". Journal of the Royal Statistical Society, Series B, Vo.67.

[11] Samiran G., (2007), "Adaptive Elastic Net: An Improvement of Elastic Net to achieve Oracle Properties", http://www.math.iupui.edu/research/preprint/2007/p07-01.pdf.

[12] Anita J., Van d., (2007), "Prediction Accuracy and Stability of Regression with Optimal Scaling Transformations", https://www.researchgate.net/publication/28648961_PredictionAccuracy_and_Stability_of_Regressionwith_Optimal_Scaling_Transformations.

[13] Rousseeuw p. and Leroy A., (1987), "Robust Regression and Outlier Detection", New York: John Wiley.

[14] Ellis, S., Morgenthaler, S., (1992), "Leverage and break down in L1 regression", Journal of the American Statistical Association Vo. 87.

[15] Chatterjee, S. and Hadi, A.S. (1988), Sensitivity Analysis in Linear Regression, Wiley, NewYork.

[16] Rousseeuw p. and Hubert M., (1997), "Recent development in PROGRESS", http://win-www.uia.ac.be/u/statis.

[17] Zou H., Zhang H. (2009), "On the adaptive elastic net with a diverging number of parameters", Institute of Mathematical Statistics, Vo.37, No.4.

[18] Wang, H., Li, G., Jiang, G., (2007), "Robust regression shrinkage and consistent variable selection through the LAD-Lasso", Journal of Business & Economic Statistics, Vo.25.

[19] Hong, D., Zhang, F., (2010), "Weighted elastic net model for mass spectrometry imaging processing. Mathematical modeling in the medical sciences", Vo.5, No.3.

[20] Li, J.-T. , Jia, Y.-M., (2010), "An Improved Elastic Net for Cancer Classification and Gene Selection", Acta Automatica Sinica, Vo.36, No.7.

[21] Wang, H., Leng, C., (2007), "Unified lasso estimation by least squares approximation", Journal of the American Statistical Association, Vo.102.

[22] Pollard, D. (1991), "Asymptotics for least absolute deviation regression estimators", Econometric Theory, Vo.7.

[23] Knight' K., (1998), "Limiting Distributions for L1 Regression Estimators under General Conditions. Annals of Statistics", Vo. 26, No.2.