مجلة جامعة ذي قار للبحوث الزراعية ، المجلد ١، العدد (٢) ٢٠١٢

استعمال أنواع الانحدار الرصين في معالجة القيم الشاذة في الانحدار الخطي البسيط

م . علي لطيف عارف
كلية العلوم
جامعة ذي قار

الخلاصة

في تحليل الانحدار بوجود متغير مستقل واحد تظهر أحيانا مشكلة وجود نقاط متطرفة تمتلك بواقي عالية مقارنة مع بواقي المشاهدات حيث تمثل قيم شاذة في مجموعة المشاهدات . تستخدم عادة طريقة المربعات الصغرى في تقدير معالم النموذج وان تحليل الانحدار يبدأ بالرسوم البيانية للبواقي مقابل المتغير المستقل وكذلك مقابل القيمة التقديرية ل y للتحقق من فرضيات النموذج فكان تحليل الانحدار الرصين بديلا عن طريقة المربعات الصغرى لوجود القيم الشاذة . تناول البحث أربعة طرائق للتقدير تم تطبيقها على مثال واحد مع استخدام معياري (AIC , BIC) في مطابقة النموذج. استخدم البرنامج الإحصائي (SAS 9.1) في تحليل النتائج

Use types of robust regression in treatment of the outliers in simple linear regression

Ali .L. Aref
College of science
Thi-Qar University

Summary

In the analysis of the simple linear regression there is only one independent variable . My be there exist a problem because there are extreme points having higher remains (Residuals) in comparison with those of observations , for there are odd values (outliers) in the groups of the observations . Usually least square method are used so as to estimation the parameters of a model .The analysis of this regressions begins with data designs of those which are remains in the opposite of an independent variety ; also , in the opposite of estimated value of Y to investigation of assumptions of that model so , the robust regression analysis in place of the least square method with outliers . This research deals with four types of assessment applied to one example along with the use of the criterions (AIC and BIC) applied through goodness fit . I used statistical program (SAS 9.1) in analyzing the results.

Introduction

One of the most important statistical tools is linear regression analysis for many fields. Nearly all regression analysis relies on method of least squares for estimation of the parameters in the method .there are several assumptions that have to be fulfilled for the ordinary least squares regression model to be valid. One of the basic assumptions of regression analysis is equality of the error variance along the predicted line , a condition called homoskedasticity. Another form of violation resides in the lack of independence of observations. This can manifest itself in term of residual autocorrelation , which can further bias the estimation of the significance tests as the error variance becomes artificially compressed by

residual correlation , When this happens , the R^2 , F and t values become inflated. Failures of these assumptions can predispose output toward false statistical significance . Another assumption is a normality of the residuals .

When there are violations of the assumption of normality of the residuals in ordinary least square regression analysis the estimation of significance becomes impaired . The regression model does not meet the fundamental assumptions , the prediction and estimation of the model my be become biased, residuals differences between the values predicted by the model and the real data , that are very large can seriously distort the prediction. When these residuals are extremely large , they are called (outliers) .

Robust regression analysis provides an alternative to a least square method and resistant
(stable) results in the presence of outliers and limits the influence of outliers . Historically , three classes of problems have been addressed techniques :

1. **Problem with outliers in the y-direction (response variable)**
2. **Problem with multivariate outliers in the covariate space (outliers in the x-space , are also referred to as leverage points)**
3. **Problem with outliers in both the y-direction and x-space .**

Many methods have been developed for these problems. Before attempting a regression analysis , the researcher should run a sample size test to be sure that will have enough statistical power to test his hypotheses . outliers : One problem with least squares occurs when there are one or more large deviations, i.e. cases whose values differ substantially from the other observations. The slope and intercept of the least squares line is very sensitive to data points which lie far from the true regression line. These points are called *outliers*, i.e. extreme values of observed variables that can distort estimates of regression coefficients, residuals plots can also be used to detect outliers , values of y that appear to be in disagreement with the model , since almost all values of y should lie with in 3σ of E(y) the mean value of y , we would expect most of them to lie with 3*S* of $\hat{y}$ If a residual is large than 3*S* (in absolute value) , we consider it an outlier and seek background in formation that might explain the reason its large value

Types of Robust Regression

There is a family of robust regression analysis that replaces the sum of squared errors as the criterion to be minimized with one less influenced by outliers.

1- *M-estimation* was introduced by Huber (1973),and it is the simplest approach both computationally and theoretically. Although it is not robust with respect to leverage points, it is still

used extensively in analyzing data for which it can be assumed that the contamination is mainly in the response direction. Consider the linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i \quad \text{---------- (1)}$$

$$= x'\beta + \varepsilon_i \quad \text{---------- (2)}$$

Where x denoted an $n \times p$ **matrix ,** β **an unknown** p **–vector of parameters**

For the ith of n observations . the fitted model is

$$y_i = a + bx_{i1} + ... + b_k x_{ik} + e_i \quad \text{---------- (3)}$$

$$= x'b + e_i \quad \text{---------- (4)}$$

Where b **is an estimates of the coefficients of regression ,The general** M**-estimator is to minimizes the** ***objective function***

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x'b) \quad \text{---------- (5)}$$

Where the function ρ **gives the contribution of each residual to the object function**

Let $\psi = \rho'$ **be the derivative of** ρ **. Differentiating the object function with respect to the coefficients , b , and setting the partial derivatives to 0 , produces a system of (k+1) estimating equations for the coefficients**

$$\sum_{i=1}^{n} \psi(y_i - x'b)x_i' = 0 \quad \text{---------- (6)}$$

Define the weight function $w(e) = \psi(e)/e$ ---------- **(7)**

And let $w_i = w(e_i)$ ---------- **(8)**

then the estimating equations may be written as

$$\sum_{i=1}^{n} w_i(y_i - x'b)x_i' = 0 \quad \text{---------- (9)}$$

Applied :- **The following data are concerned with television rating where y is represent splay news rating of television data and x is lead rating. (polyu.edu.hk/ mathwong/ch2)**

***Table 1* :**

X	Y	x	y
2.50	3.80	5.50	4.35
2.70	4.10	5.70	4.15
2.90	5.80	5.90	4.85
3.10	4.80	6.10	6.20
3.30	5.70	6.30	3.80
3.50	4.40	6.50	7.00
3.70	4.80	6.70	5.40
3.90	3.60	6.90	6.10
4.10	5.50	7.10	6.50
4.30	4.15	7.30	6.10
4.50	5.80	7.50	4.75
4.70	3.80	2.50	1.00
4.90	4.75	2.70	1.20
5.10	3.90	7.30	9.50
5.30	6.20	7.50	9.00

Television rating data

The OLS analysis of table 2 indicates that x have a significant influence on y at the 5% level

Figure (1): The four observations are outliers

Table 2: OLS for data in table 1

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	**1.70654**	0.81715	2.09	0.0460
x	**1**	**0.66536**	0.15521	4.29	0.0002

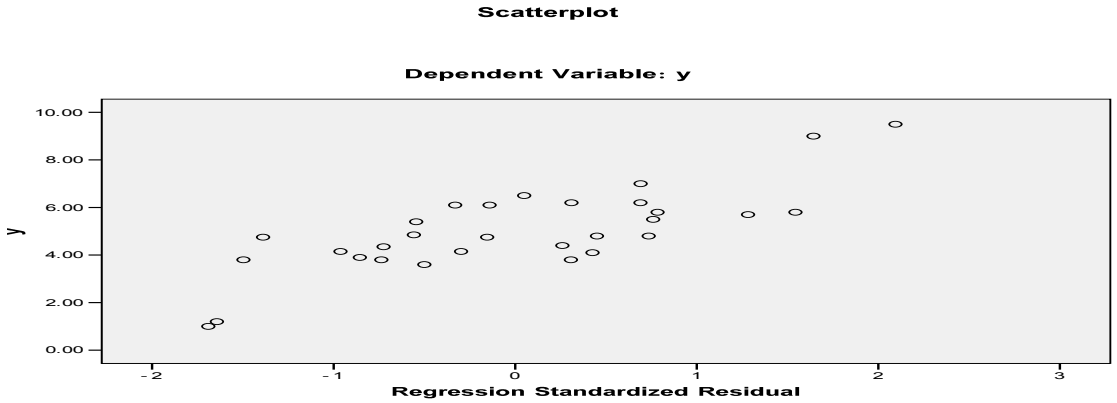

Table 3 displays model information and summary statistics for variable in the model , the robust analysis also indicates x has significant impact on y

Table 3 : Results of ***M-Estimation***

Model Information							
Data Set							
WORK.ALI							
Dependent Variable							
y							
Number of Independent Variables							
1							
Number of Observations							
30							
Method							
M- Estimation							
Number of Observations Read							
30							
Number of Observations Used							
30							
Parameter Estimates							
			Standard	95% Confidence		Chi-	
	Parameter	DF	Estimate	Error	Limits	Square	Pr > ChiSq
Intercept	1	1.9748	0.8583	0.2925	3.6571	5.29	0.0214
x	1	0.6084	0.1630	0.2889	0.9280	13.93	0.0002
Scale	1	1.4616					

2 - *Least Trimmed Square (LTS)* estimation is a high breakdown value method introduced by Rousseeuw (1984). The breakdown value is a measure of the proportion of contamination that a procedure can withstand and still maintain its robustness. The performance of this method was improved by the FAST-LTS algorithm of Rousseeuw and Van Driessen (1998).

One way to eliminate possible outliers is to run the analysis on trimmed or winsorized distributions. Distributions that have their outliers trimmed prior to the analysis are sometimes called trimmed means procedures. While trimming a distribution means truncating it at the *t* and 1- *t* quantile (Wilcox, 1997), a distribution means setting the values at or more extreme than the *t* quantile to that of the *t* quantile on one tail and setting those values at or more extreme than the 1- *t* quantile to those of that quantile. The trimmed or winsorized distribution is then estimated by minimizing the sum of squared absolute residuals. By trimming the alpha rejection region, the distorting effects of influential outliers could be pruned from the variables prior to processing

Least Trimmed Square (L T S) = $\min \sum_{i=1}^{q} e_i^{\,2} = \sum_{i=1}^{q} |y_i - x_i b|^2$ **---------- (10)**

Where $q = [n + p + 1]/2$ **is the number of observations included in the calculation of the estimator**

The range of q is (n/2)+1 ≤ q ≤ (3n+ p+1) / 4

And n = sample size , p = number of parameters

Table 4 :output of *LTS - Estimation*

Model Information			
Data Set			
WORK.ALI			
Dependent Variable			
y			
Number of Independent Variables			
1			
Number of Observations			
30			
Method			
LTS Estimation			
LTS Parameter Estimates			
Parameter		DF	Estimate
Intercept		1	3.4177
x		1	0.3070
Scale (sLTS)		0	1.2311
Scale (Wscale)		0	1.3736
Diagnostics Summary			
Observation			
Type	Proportion	Cutoff	
Outlier	0.0000	3.0000	

3- *MM – Estimation* Introduced by Yohai (1987) combines high breakdown value estimation and M estimation . Its has both the high breakdown property and a higher statistical efficiency than S estimation. Its has three steps :

1 - Compute an initial (consistent) high breakdown value estimation $\hat{\theta}'$.

2 - Find $\hat{\sigma}'$ such as

$\sum_{i=1}^{n} \chi(\frac{y_i - x_i^T \hat{\theta}'}{\hat{\sigma}'}) = (n-p)\beta$ **or** $\sum_{i=1}^{n} \left(\frac{y_i - x_i b}{c_0 s}\right) = (n-p)\beta$ **,** ---------- **(11)**

$c_0 =$ **tuning constant =1.548**

Where $\beta = \int \chi(s) d\phi(s).$ **=0.5**

Robust initial regression coefficients are used as starting values and found by minimizing a scale parameter, S , while χ may be one of the several bounded loss functions that serves the purpose of minimizing the empirical influence the troublesome residuals . χ is an integral of $\chi(s)$.

Table 5 : Result of MM estimation

Model Information							
Data Set	WORK.ALI						
Dependent Variable	y						
Number of Independent Variables	1						
Number of Observations	30						
Method	MM Estimation						
Parameter Estimates							
Parameter	DF	Estimate	Standar Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.4276	0.9255	0.6137	4.2416	6.88	0.0087
x	1	0.5137	0.1772	0.1665	0.8609	8.41	0.0037
Scale	0	1.4297					

4 - ***S- Estimation*** **is a high breakdown value method introduced by Rousseeuw and Yohai (1984.) With the same breakdown value, it has a higher statistical efficiency than LTS estimation. is defined as the p-vector**

$\hat{\theta}_s = \arg\min_\theta S(\theta)$ **where the dispersion S(θ) is the solution of**

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - x_i^T\theta}{s}\right) = (n-p)\beta \quad \text{---------- (12)}$$

β **is set to** $\int \chi(s)d\phi(s).$ **such that** $\hat{\theta}_s$ **and** $s(\hat{\theta}_s)$ **are asymptotically consistent estimates of θ and S for the Gaussian regression model. The breakdown value of the S estimate is**

$$\frac{\beta}{\sup_s \chi(s)} \quad \text{---------- (13)}$$

Table 6 :result of *S-Estimation*

Data Set	WORK.ALI						
Dependent Variable	y						
Number of Independent Variabl	1						
Number of Observations	30						
Method	S Estimation						
Parameter Estimates							
Parameter	DF	Estimate	mStandard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.9913	0.9368	1.1552	4.8274	10.20	0.0014
x	1	0.3980	0.1800	0.0452	0.7508	4.89	0.0270
Scale	0	1.4224					

Robust Inference
Goodness of Fit

The robust version of R^2 is defined as

$$R^2 = \frac{\sum \rho(\frac{y_i - \hat{\mu}}{\hat{s}}) - \sum \rho(\frac{y_i - x_i^T \hat{\theta}}{\hat{s}})}{\sum \rho(\frac{y_i - \hat{\mu}}{\hat{s}})} \quad \text{----------(14)}$$

And the robust deviance is defined as the optimal value of the object function on the σ^2-scale

$$D = 2(\hat{s})^2 \sum \rho(\frac{y_i - x_i^T \hat{\theta}}{\hat{s}}) \quad \text{---------- (15)}$$

Where ρ is the objective function for the robust estimate , $\hat{\mu}$ is the robust location estimator , and $\hat{s}$ is the robust scale estimator in the full model .
The information Criterion is a powerful tool for model selection . The counter part of the Akaike (1974) AIC criterion for the robust regression is defined as

$$AICR = 2\sum_{i=1}^{n} \rho(r_{i:p}) + \alpha p \quad \text{---------- (16)}$$

Where $r_{i:p} = (y_i - x_i^T \hat{\theta}) / \hat{\sigma}$, $\hat{\sigma}$ is some robust estimate of σ, and $\hat{\theta}$ is the robust estimator of θ with a p-dimensional design matrix . As in AIC , α is the weight of the penalty for dimensions .
Also therese another measure using in the robust residuals is BIC is defined as

$$BICR = 2\sum_{i=1}^{n} \rho(r_{i:p}) + p\log(n) \quad \text{---------- (17)}$$

Table 7 : the following goodness-of-fit table

Goodness-of-Fit	
Statistic	Value
R-Square	0.2784
AICR	26.2745
BICR	30.1479
Deviance	49.8733

Conclusion:

1- Robust regression analysis provides an alternative to a least squares regression model when
fundamental assumptions are unfulfilled by the nature of the data .
2- There is a many types of robust regression diagnostic tool and appropriate for that particular
application can be found that replaces the sum of squared errors as the criterion to be minimized
with one less influenced by outliers
3- Even if robust regression is better suited for those who do not want to put much effort into
testing the assumptions, it is so far difficult to use
4- Robust methods does not protect against problems that are due to curvilinear or non-linear
models, heteroscedasticity, and autocorrelation

References

1 - Chen , Colin (2002).Robust regression and outlier detection with the ROBUSTREG procedure
.SUGI paper 265-27 .SAS Institute :Cary , NC.
2- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics:
The Approach Based on In5uence Functions. Wiley, New York
3- Hertier, S., Ronchetti, E., 1994. Robust bounded-in5uence tests in general parametric models. J.
Amer. Statist. Assoc. 89, 897–904.
4- Huber, P.J., 1975. Robustness and designs. In: Srivastava, J.N. (Ed.), A Survey of Statistical
Design and Linear Models. North- Holland, Amsterdam, pp. 287–303.
5- Meer, P., Mintz, A. & Rosenfeld, A. (1991) Robust regression methods for computer vision: a
review. International Journal of Computer Vision , 6: 59–70.
6- Rousseeuw, P. J. & Leroy, A. M. (1987) Robust Regression and Outlier Detection. New York:
John Wiley and Sons Inc.
7- Zaman, A., Rousseeuw, P.J., Orhan, M. (2001), "Econometric applications of high-breakdown
robust regression techniques", Econometrics Letters, 71, 1-8.