Regression Depth for Statistical Depth Function

Mohammed Al-Guraibawi [1,a)] and Hadeel Habib Kamel [Name2, 3, b)]

[1]Al-Furat Al-Awsat Technical University- Diwaniyah Technical Institute, Iraq
[2]Al-Qadisiyah University- College of Administration and Economics, Iraq

[a)]dw.moh2@atu.edu.iq
[b)]hadeel_8080@gmail.com

ABSTRACT

The statistical depth function is one of the modern approaches that can be used for developing multivariate robust regression based on robust estimates of the location and dispersion matrix. One merit advantage of the depth concept is that it can be used directly to provide deeper estimation functions for data location and regression parameters in a multidimensional environment. The deeper estimation functions induced by depth are expected to inherit the desired and inherent robustness properties (such as limited maximum bias, impact function, and high breaking point) as do their counterparts at univariate sites. Investigation. The main objective of this article is to check the power of the statistical depth function throw the depth regression, it turns out that the deepest functional projection possesses a finite effect function and the best possible asymptotic breakpoint as well as the best breaking point of a finite sample compared with some classical and robust existed method.

1. INTRODUCTION

For the multivariate linear regression model:

$$y = X\beta + \varepsilon \ldots \quad (1)$$

If n is a sample size and p is a number of predictors, then; $y(n \times 1)$ and $\beta\ (p \times 1)$ are the vectors of response and unknown regression coefficients, respectively. $X\ (n \times p)$ is a matrix of explanatory variables and $\varepsilon(n \times 1) \sim IID(0, \sum_e)$ is the random error vector distributed as identical normal distribution with zero mean and constant variance ([3],[8]).

Linear regression model is widely used in many applications in statistics. It is simple to compute classical least squares regression, which reduces the sum of the squares of the residual but it destroys when one or more unusual points exist in the dataset. The robust methods are alternatives approaches which often computationally intensive. Following we present some common robust regression techniques that have well combinatorial algorithms.

1- M-estimator

the M-estimators is proposed by Huber ([4], [8]) . The class of M-estimators can be considered as a generalization of maximum-likelihood estimator. The M-estimator is determined by minimizing the sum of a less rapidly increasing function of residuals, as follows

$$\min_{\beta} \sum_{i=1}^{n} \rho(r_i) = \min_{\beta} \sum_{i=1}^{n} \rho\left(y_i - \sum_{i=1}^{n} x_{ij}\hat{\beta}_j\right) \ldots \qquad (2)$$

where ρ is a particular function determines the contributions of each residuals in the objective function

1- Least Median of Squares Regression

least median of squares (LMS) as an alternative robust technique proposed by Rousseeuw (1984) ([4],[8]). The LMS estimator is obtained by reducing the median (Med) of squared errors, as

$$\min Med\ (e_i^2) = \min Med \left(y_i - \sum x_{ij}\hat{\beta}_j\right)^2 \ \ldots \quad (3)$$

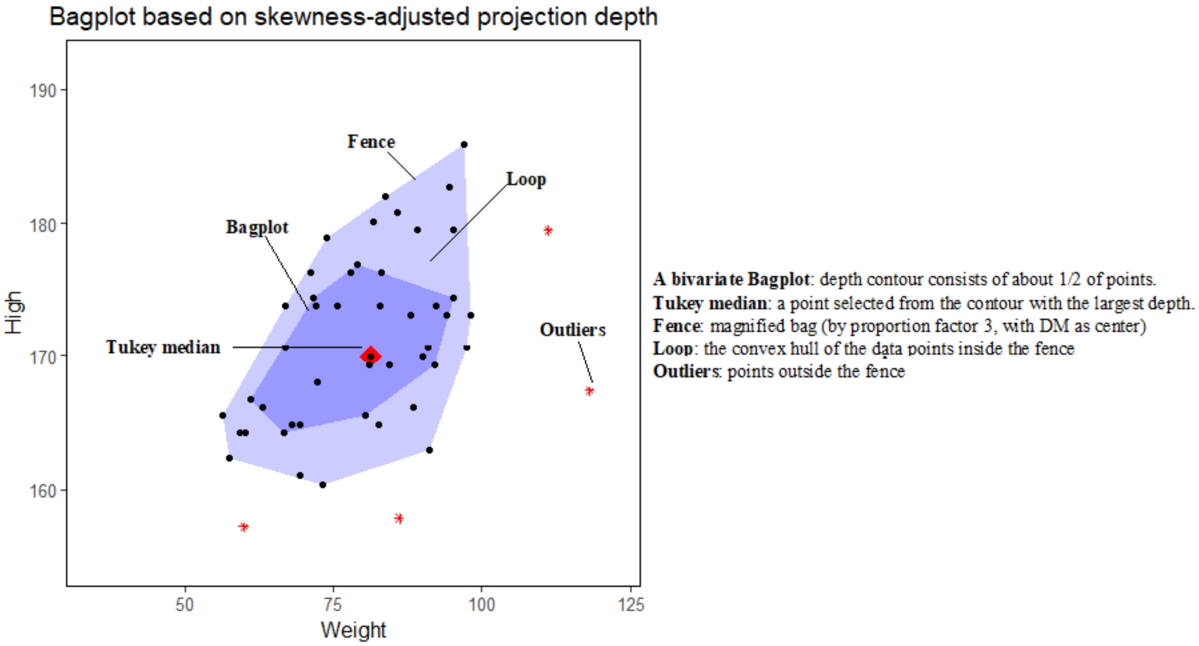
The LMS estimator has high breakdown point of 50 %, but it has low relative efficiency of 37%. (see [4], ,[6], [8], [9]) in this study, we discuss the concept of statistical depth and present some applications of depth statistics such as median depth, bagplot, and regression depth. some examples are given to investigate and compare the depth statistics application with some classical and robust existing method. The rest of the article is organized as follows:

in Section 2, the bagplot of the statistical depth function is briefly discussed with some figures. In section 3, the regressions depth is illustrated with some depth statistics properties. Section 4 presents the regression depth briefly. Section 5, view example of applying regression depth with discussion. Finally, some conclusions are given in Section 6.

2. Bagplot of The Statistical Depth Function

The statistical depth function (SDF) is an approach suggested by John Tukey (1975) ([1], [12]). The SDF is determined how close an arbitrary point of the space has existed to an implicitly specified location of a data cloud. In addition, Tukey introduced a "depth median" (DM) which is the 'deepest' point in a specific data cloud (Tukey, 1975). The DM is the deepest point which is enclitic by a "bag" containing the half points with the largest depth. There are a lot of subjects like economics, social sciences that cannot be modeled easily, due to our knowledge of economic rules is not sufficient for effective parametric modeling or the datasets containing outliers or missing data, hence the SDF is the effective approach to deal with it.

In the SDF, the bagplot is a modified shape for the well-known boxplot proposed by Rousseeuw, Ruts, and Tukey ([10], [11]). In the bivariate case, the box of the boxplot changes to a convex hull, the bag of bagplot. As shown in figure 1, in the bag, there is 50 percent of all observations. The fence separates observations inside and outside the fence. The loop is stated as the convex polygon that holds all observations inside the fence if all points are on a straight line you get a classical boxplot. increasing the bag by a proportion of 3 results in the "fence" as shown in figure 1. points between the bag and the fence are flagged by a light gray loop, whereas points outside the fence are identified as outliers. The bagplot conceive the location, dispersion, overlapping, skewness, and tails of the dataset [11].



A bivariate Bagplot: depth contour consists of about 1/2 of points.
Tukey median: a point selected from the contour with the largest depth.
Fence: magnified bag (by proportion factor 3, with DM as center)
Loop: the convex hull of the data points inside the fence
Outliers: points outside the fence

FIGURE 1: Bagplot based on adjusted projection depth [12]

3. Data Depth

The data depth of multivariate points is present by univariate (individual) number defined by a depth function. That value of depth is used to identify outliers in the data. A depth function is a real-valued function that produces a “center-outward” ordering of the multivariate data. Recently, many depth function has been suggested in the literature ([1], [2], [13], [14]).

- Tukey (1975) halfspace
- The simplicial depth (Liu, 1990)
- The majority depth (Singh,1991; Liu and Singh, 1993),
- The projection depth (Liu, 1992; Zuo, 2003),
- The Mahalanobis depth (Liu and Singh, 1993; based on Mahalanobis, 1936)
- The spatial depth (Serfling, 2002; based on Chaudhuri, 1996),
- The elliptical depth (Elmore, 2005),
- The spherical depth (Elmore, Hettmansperger and Xuan. 2006),
- And recently the triangle depth (Liu and Modarres, 2010).

The depth data has desirable statistical properties due to it depend on the depth function D(x; p). The depth function D(x; p) has the following properties:-

1- Affine invariant, $D(x;P) = D(Ax+b;\ PAx+b)$ for every nonsingular matrix $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^{d}$.
2- Vanishes at infinity: $D(x;P) \Longrightarrow 0 \quad if\ \|x\| \Longrightarrow 0$.
3- Upper semicontinuity: $\{x\ \in \mathbb{R}^{d}: D(x;P) \geq \alpha\}$is closed.
4- Monotonicity relative to deepest point:

The classical estimation methods are extremely sensitive to outlying observations. A popular approach to deal with this issue is to use robust variance-covariance matrix. One of the common practice to obtain robust estimators is to use concept of depth statistics [13] Applying depth in construct for getting such estimator is easy due to depth has “center outward ordering”. The depth has ability to increase at the center of the data cloud and minimize along all direction for that center [1]. Observations those extreme with respect to the bulk of data will be down weighted by depth.

The value of the depth is proportional “inversely” to the distance from the center for the data cloud, as the data is closer to the center the greater its depth. On the contrary, the lower the depth value, the further away from the center it is. From the foregoing, it can be seen that the depth of a point can shift to its outlyingness (and vice versa). It will be concluded from the above that the process of converting the depths of points in the data set into outlyingnesses, will help us to identify points with far outliers on the basis of specific cut-off point [12].

4. Robust Regression Depth

The least-squares approach (LS) is the usual method of getting estimators of linear regression. However, it is well-known that LS estimators are highly sensitive to unusual points. A popular method alternative to LS is thus to use robust estimators of location μ_x and scatter Σ_x. In robust regression approach, many variance covariance matrix has been suggested such as MVE and MCD by Rousseeuw ([8], [9]). The regression depth (RD) is a quality measure for robust linear regression. As a statistical vision, "the RD of a hyperplane (RHP) is the minimum number of residuals that need to change the sign to be non-fit".

For k-dimensional independent vector $X = (X_1, X_2, \dots, X_k)'$ and m-dimensional $Y = (Y_1, Y_2, \dots, Y_k)$ for $K \geq 1$ and $m \geq 1$. The multiple regression model is:

$$Y = \alpha + \beta' X + \epsilon$$

where β is the $(k \times m)$ coefficient matrix and α is the m-dimensional constant term vector. The error term ϵ is identically distributed with mean zero and covariance Σ_e. Let $Z = (Y', X')$ and the location and variance matrix of Z given as μ_Z and Σ_Z, respectively. The partition of location and scatter of Z is as follows [2]:

$$\mu_Z = (\mu_Y', \mu_X') \text{ and } \Sigma_Z = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} \quad \dots \quad (4)$$

with $\Sigma_{XY} = \Sigma_{YX}{}'$,

and, $z_i = (y_i', x_i')'$ where y_i and x_i $(1 \leq i \leq n)$ are the response vector and predicted vector, respectively. and, $z_i = (y_i', x_i')'$ where y_i and x_i $(1 \leq i \leq n)$ are the response vector and predicted vector, respectively. Assume $\hat{\Sigma}_{XX}$ to be invertible and suppose $\hat{\mu}_Z$ and $\hat{\Sigma}_Z$ are the estimates of μ_x and Σ_x , respectively. The estimators of α , β and Σ_e are as follows)[2], [14]):

$$\hat{\alpha} = \hat{\mu}_Z - \hat{\beta}$$

$$\hat{\beta} = \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY} \quad \dots . \quad (5)$$

$$\hat{\Sigma}_e = \hat{\Sigma}_{YY} - \hat{\beta}'\hat{\Sigma}_{XX}\hat{\beta}$$

In RD, robust estimates of μ_x and Σ_x are used in depth based estimators (see Serfling, 2006; Zuo, Cui and He, 2004; Zuo, Cui and Young, 2004) given as:

$$\hat{\mu}_Z = \frac{\sum_{j=1}^{n} w_1\{\gamma(Z_j)\}Z_j}{\sum_{j=1}^{n} w_1\{\gamma(Z_j)\}} \quad .... \quad (6)$$

$$\hat{\Sigma}_Z = \frac{\sum_{j=1}^{n} w_2\{\gamma(Z_j)\}(Z_j - \hat{\mu}_Z)(Z_j - \hat{\mu}_Z)'}{\sum_{j=1}^{n} w_2\{\gamma(Z_j)\}} \quad .... \quad (7)$$

where: $\gamma(Z_j)$ represent the depth of Z_j, $1 \leq j \leq n$.

w_1 and w_2 are not non-decreasing and non-negative weight functions (WF) and they are not imposed to be the same. The weight function w_i, is given as follows:

$$w_i(v) = \frac{exp\left[-p\left\{1 - (\frac{v}{c})^{2i}\right\}^{2i}\right] - \exp(-p)}{1 - \exp(-p)} I(0 < v < \theta) + I(\theta < v < 1) \ldots (8)$$

Where $I(.)$ is the indicator function, $(0 < \theta < 1)$ and $p > 0$ for $i = 1, 2$. p control the grade of approximation. According to Zuo et. al (2004), a consistent estimate of v is imposed on the median of the depth values and p is assumed to be 100. The WF assigns weight "one" to half of the observations with larger depth and this equipoise efficiency with robustness. The second half of the observations with smaller depth may be assigned as outliers, so they give lower weights. Many WF can be used that satisfy appropriate properties [14].

5. Example and discussion

In this section, we consider one example investigate targets of this study and to compare the performance of robust regression depth versus other existing methods. The first example is the well-known Hawkin Bradu Kass's Artificial Dataset (HBK). The HBK dataset was introduced by Hawkins et al. (1984) [8]. The data has four variables, one is the response variable and the rest are the independent variables. The dataset has 75 cases with 14 outliers (cases 1-14) [3]. Table 1 present the depth values for the points of the dataset. It's clear to see that observations numbered (1-14) have the lowest depth values, so they identify as outliers. On the other hand, we observed from Figure 2 (a, b, and c) that the points (1-14) lie outside of the bagplot indicating that they are outliers. Figure 2-d shows the bagplot for the data, where we find that the fence of the bagplot expands greatly towards the anomalous data, while we find the outliers are located outside the bagplot

Figure 3 shows the fit of the regression line for the HBK dataset, where we find that the regression line of the least-squares method (LS) is far away from the data cloud, while we find that the regression lines of the robust methods are all passed from the bulk of dataset.

Table 1: depth values for Hawkin Bradu Kass's Artificial Dataset

Depth values									
1	0.0352	16	0.4696	31	0.5088	46	0.3306	61	0.3390
2	0.0337	17	0.5472	32	0.4550	47	0.2558	62	0.2596
3	0.0336	18	0.5675	33	0.3139	48	0.5445	63	0.4341
4	0.0337	19	0.5417	34	0.2642	49	0.3608	64	0.3422
5	0.0333	20	0.3776	35	0.5605	50	0.3896	65	0.3016
6	0.0339	21	0.3804	36	0.3289	51	0.3610	66	0.3540
7	0.0322	22	0.4254	37	0.3232	52	0.3294	67	0.4062
8	0.0341	23	0.3584	38	0.3001	53	0.2674	68	0.2535
9	0.0343	24	0.4267	39	0.2642	54	0.3623	69	0.3553
10	0.0347	25	0.2901	40	0.4732	55	0.6538	70	0.4378
11	0.0427	26	0.2693	41	0.5661	56	0.4694	71	0.5921
12	0.0448	27	0.3567	42	0.3458	57	0.2886	72	0.6287
13	0.0390	28	0.3596	43	0.3505	58	0.3836	73	0.4654
14	0.0296	29	0.3712	44	0.2996	59	0.5686	74	0.3485
15	0.3760	30	0.3647	45	0.2867	60	0.3511	75	0.3315

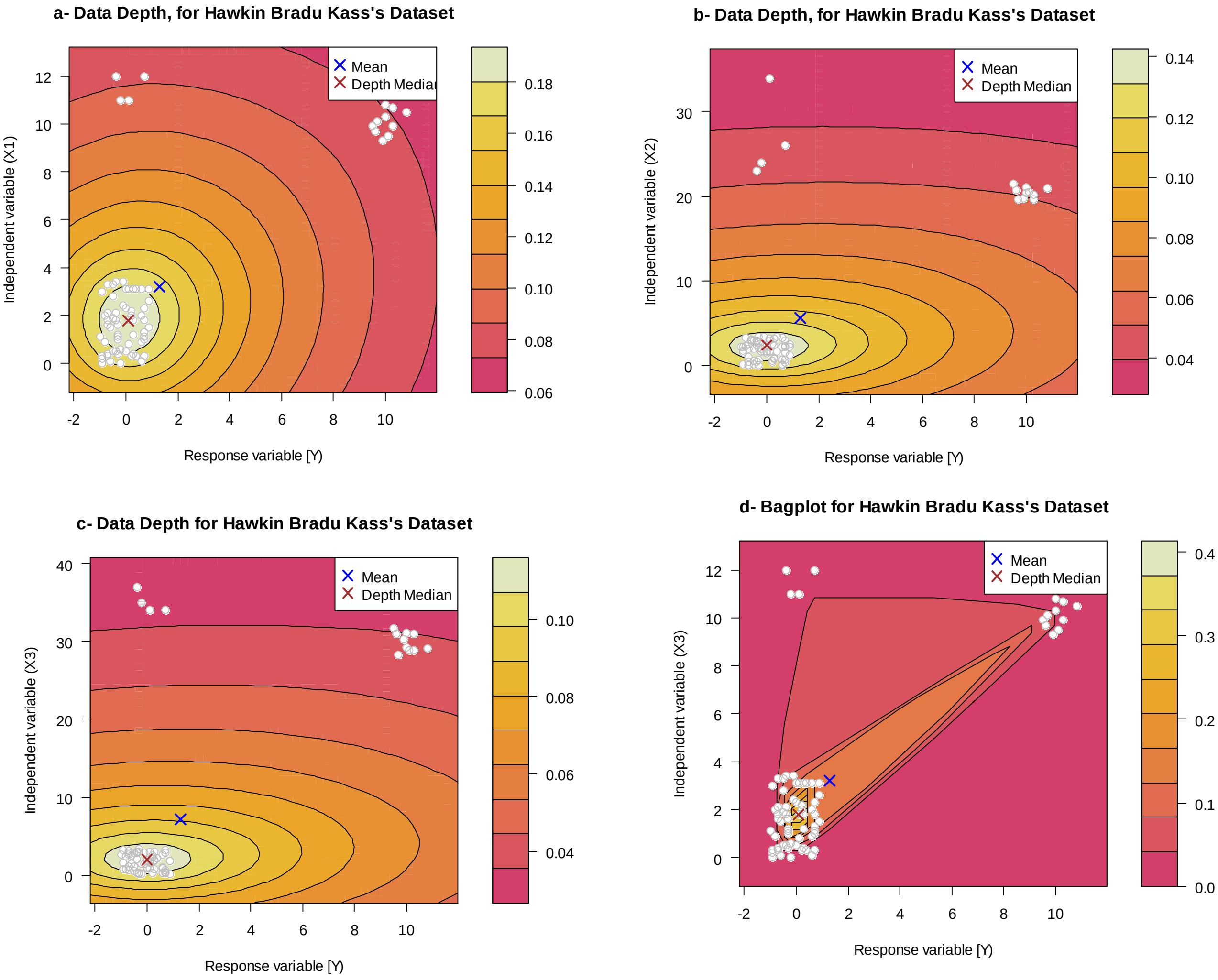


FIGURE **2:** a, b and c represent plots for independent variables (x1, x2 and x3) v.s response variable for HBK dataset. “d” represents the pagplot for x1 v.s y [12]

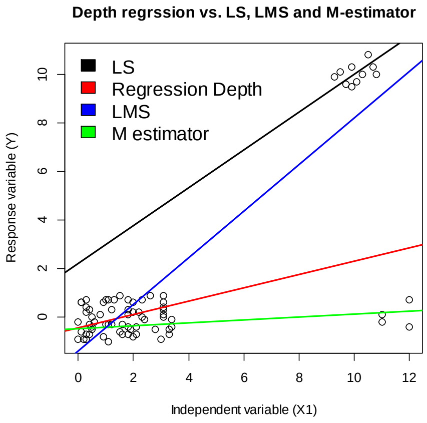


FIGURE 3: Regression lines for LS, LMS, M-estimator and Regression depth [12]

5. Conclusions

In this study, we discussed the concept of depth statistical function, which is characterized by many good merits, such as (Affine invariant, Vanishes at infinity, and Upper semicontinuity). The box plot was also discussed as to how to use it in diagnosing outliers in the data. The pagplot is dependent on the Tukey Median which has great property in representing the data, unlike the arithmetic mean, which was far away from the data cloud significantly. Regression Depth was also discussed as one of the robust methods for representing the regression line for the data, as is the case in other strong methods such as the LMS method and the mother method. The results of the example confirmed that the concept of depth statistics has good merits in identifying outliers and fitting of the regression line in the presence of outliers.

6. References

1. Dovoedo, Yinaze Herve. *Contributions to outlier detection methods: Some theory and applications*. The University of Alabama, 2011.
2. Dutta, Subhajit, and Marc G. Genton. "Depth-weighted robust multivariate regression with application to sparse data." *Canadian Journal of Statistics* 45.2 (2017): 164-184. p.p 1-4.
3. Habshah, M., M. R. Norazan, and A. H. M. Rahmatullah Imon. "The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression." *Journal of Applied Statistics* 36.5 (2009): 507-520.
4. Farcomeni, Alessio, and Luca Greco. *Robust methods for data reduction*. CRC press, 2016.
5. Grentzelos, Christos, Chrysseis Caroni, and Inmaculada Barranco-Chamorro. "A comparative study of methods to handle outliers in multivariate data analysis." Computational and Mathematical Methods 3.3 (2021): e1129, p.p 2-3.
6. Maronna, Ricardo A., et al. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.
7. Mosler, Karl. "Depth statistics." *Robustness and complex data structures*. Springer, Berlin, Heidelberg, 2013. 17-34.
8. Rousseeuw, Peter J. "Least median of squares regression." *Journal of the American statistical association* 79.388 (1984): 871-880.
9. Rousseeuw, P. J., and A. M. Leroy. "Robust Regression and Outlier Detection. New York: John Wiley& Sons." (1987).
10. Rousseeuw, Peter J., and Ida Ruts. "The depth function of a population distribution." *Metrika* 49.3 (1999): 213-244
11. Rousseeuw, Peter J., and Mia Hubert. "Computation of robust statistics: depth, median, and related measures." *Handbook of discrete and computational geometry*. Chapman and Hall/CRC, 2017. 1541-1554.
12. Pokotylo, Oleksii, Pavlo Mozharovskyi, and Rainer Dyckerhoff. "Depth and depth-based classification with R-package ddalpha." *arXiv preprint arXiv:1608.04109* (2016).
13. Serfling, Robert. "Depth functions in nonparametric multivariate inference." *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 72 (2006).
14. Zuo, Yijun, and Robert Serfling. "General notions of statistical depth function." *Annals of statistics* (2000): 461-482.