Robust Estimation for Independent Non-Homogeneous Observations Using Density Power Divergence With Applications to Linear Regression

Prithwjit Ghosh(211349)
Gagan Deep(211304)
Sarvesh Singh Khushwaha(211455)
Gaurav(211303)

Submitted to
Dr. Arnab Hazra

November 14, 2022

Contents

1 Abstract

The theoretical estimation uses the MLE method heavily due to its high efficiency but it has also a drawback as it is highly non-robust. So, in this paper, we build a bridge between robustness and efficiency via a different divergence measure. As those estimators depend on a single tuning parameter alpha and changing the values of alpha we can convert our estimators from efficient to robust.

2 Introduction

In the statistical Inference problem, or more specifically, in the estimation problem, we have to estimate the unknown parameter (for the parametric inference setup) and the completely unknown distribution (for the non-parametric setup). Now based on different distributions and needs the original theory becomes too much revolved into the hardcore theory. Some of these are fascinating and some of the results of the methods are popular. But the main reason for this popularity may be

- simple to implement
- closed form expression
- more readable and explainable

Besides the above positive points, they may not have some essential points or criteria e.g. let us consider one estimator T which is unbiased for some population parameter say,θ. Now the variance of the estimator, which is also another essential criterion, may be very high, making the estimator practically inefficient to use.

Here in this paper, we focus on the effect of outliers, one of the most important criteria, to judge whether an estimator is good or not. In statistical words, this criterion is known as **Robust Estimation**. Here we mainly discuss the estimation problem where the Maximum Likelihood Estimation(MLE) procedure may not be as good as in the usual setup. But one main restriction is that here we discuss the cases where the random variables belong to the same family. Still, they are not identically distributed i.e. they are independently distributed but not identically. One of the most usual examples is the linear regression setup.

Here we first develop a general theory to estimate the parameters when they are independently distributed but not identically, based on the Density Power Divergence(DPD) estimators. Then as a particular case, we will show that we can get the estimators corresponding to the parameters in the model and also the unknown error variance for the linear regression setup(under the normality assumption of the random error). Then we will compare the Ordinary Least Square estimators to the DPD estimators for the cases when there are some extreme outliers(where the assumption of the normality may be failed).

First, we will discuss the formulation and the implementation of the DPD estimators for the non-homogeneous but independent observation. Then we will see the asymptotic behavior of this estimator. After that, we will discuss the structure for the linear regression case when the underlying distribution is normal. All the theoretical results will be developed based on the simulated data and will show that this method is working on the simulated data set. Finally, all the methods and the theory will be applied to real-world data.

3 The Density Power Divergence(DPD) estimator for the independent and non-homogeneous data

The theory of the DPD estimator was first introduced by Basu et all in 1950. But he implements the method for the iid setup. But we will follow the work of Dr. Abhik Ghosh and Dr. Ayanendranath Basu, the further extension to the case of non-homogeneous but independent data.

Let us suppose g and f be the density functions, then the density power divergence $d_\alpha(f, g)$ is defined as

$$d_\alpha(f,g) = \int \{f^{1+\alpha} - (1+\frac{1}{\alpha})f^\alpha g + \frac{1}{\alpha}g^{1+\alpha}\} \qquad for\ \alpha\ >\ 0 \tag{1}$$

$$d_0(f,g) = \int g\ln(\frac{f}{g}) \qquad for\ \alpha\ =\ 0 \tag{2}$$

here α is the tuning parameter and needs to be tuned before using the implemented model. Now $d_\alpha(f, g)$ is not defined for $\alpha = 0$, hence we consider the limiting value of $d_\alpha(f, g)$ as $\alpha \to 0$ as $d_0(f, g)$. This parameter also links two famous statistical quality measures for finding the relatively best estimators i.e. *Efficiency* and *Robustness*. If the value of α is very small say nearly equal to zero, then the model is highly efficient and if α is nearly equal to 1, it is highly robust. Therefore we have to choose a suitable value of α for which the robustness should be sufficiently high after sacrificing efficiency a little.

Let G be the proper data generating distribution, g be the corresponding densities, and f_θ be the modeling densities for different values of θ i.e. $f_\theta \in F_\theta$ where $F_\theta = \{f_\theta : \theta \in \Theta\}$. Now we need to find a value theta for which $d_\alpha(.,.)$ is minimum i.e the distance between the densities or more statistical divergence is smallest. For this, we need to consider only those terms which are dependent on θ, since we will minimize the divergence concerning theta. From equation (1) we can consider only the first two terms since g is independent of θ.

Let $X_1, X_2, ..., X_n$ be n observations or a random sample of size n from the true distribution g, then equation (1) can be written as -

$$d_\alpha(f_\theta, g) = \int f_\theta^{1+\alpha} - (1+\frac{1}{\alpha})\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i)^\alpha$$

and the estimator of $\int f_\theta^\alpha g$ is obtained by the simple Monte-Carlo simulation.

Now the original data, say $y_1, y_2, ..., y_n$ follows some distribution independently but not identically,say $g_1, g_2, ..., g_n$ respectively. Now we will compute to estimate the parameter θ by minimizing the average of the density power divergence for the density estimate of each data point and the corresponding model density. In more mathematical language,

$$\frac{1}{n}\sum_{i=1}^{n} d_\alpha(\hat{g}_i, f_i(., \theta)$$

Now the $\hat{g}_i$ can't easily be estimated by the data, since there is only one data point corresponding to each distribution $g_i\ \forall i = 1(1)n$. So we will consider that the corresponding

estimate of g_i i.e. $\hat{g_i}$ is exactly y_i. So the density power divergence is finally of the form -

$$d_\alpha(y_i, f_i) = \int f_i(y,\theta)^{1+\alpha} dy - (1 + \frac{1}{\alpha}) f_i(y_i,\theta)^\alpha + K$$

where K is a constant independent of θ The objective function,

$$H_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \int f_i(y,\theta)^{1+\alpha} dy - (1 + \frac{1}{\alpha}) f_i(y_i,\theta)^\alpha \tag{3}$$

or,

$$H_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} V_i(\theta) \tag{4}$$

So, we have to minimize it and equate it to 0

$$\nabla \sum_{i=1}^{n} V_i(\theta) = 0$$

From, equation (3), we can simplify it as

$$\frac{1}{n} \sum_{i=1}^{n} (1+\alpha) \int f_i(y,\theta)^\alpha f_i'(y,\theta) dy - (1 + \frac{1}{\alpha}) \alpha f_i(y_i,\theta)^{\alpha-1} f_i'(y_i,\theta) = 0$$

$$\implies \frac{1}{n} \sum_{i=1}^{n} (1+\alpha) \int f_i(y,\theta)^{\alpha+1} \frac{f_i'(y,\theta)}{f_i(y,\theta)} dy - (1 + \frac{1}{\alpha}) \alpha f_i(y_i,\theta)^\alpha \frac{f_i'(y_i,\theta)}{f_i(y_i,\theta)} = 0$$

$$\implies \frac{1}{n} \sum_{i=1}^{n} (1+\alpha) \int f_i(y,\theta)^{\alpha+1} \nabla ln(f_i(y,\theta)) - (1 + \frac{1}{\alpha}) \alpha f_i(y_i,\theta)^\alpha \nabla ln(f_i(y_i,\theta)) = 0$$

$$\implies \frac{1}{n} \sum_{i=1}^{n} (1+\alpha) \int f_i(y,\theta)^{\alpha+1} \nabla ln(f_i(y,\theta)) - (1 + \frac{1}{\alpha}) \alpha f_i(y_i,\theta)^\alpha \nabla ln(f_i(y_i,\theta)) = 0$$

$$\implies \sum_{i=1}^{n} (1+\alpha) \int f_i(y,\theta)^{\alpha+1} u_i(y,\theta) - (1+\alpha) f_i(y_i,\theta)^\alpha u_i(y_i,\theta) = 0$$

$$\implies \sum_{i=1}^{n} \int f_i(y,\theta)^{\alpha+1} u_i(y,\theta) - f_i(y_i,\theta)^\alpha u_i(y_i,\theta) = 0 \tag{5}$$

where, $u_i(y,\theta) = \nabla ln(f_i(y,\theta))$

Now as $\alpha \to 0$ (using (2)), we can find the corresponding divergence,

$$d_\alpha(y_i, f_i) = \frac{1}{n} \sum_{i=1}^{n} -ln(f_i(y_i,\theta)) + K$$

,where K is some constant. So, the objective function is,(after taking the derivative and equating it to 0)

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\nabla ln(f_i(y_i,\theta)) = 0$$

$$H_n(\theta) = \sum_{i=1}^{n} u_i(y_i,\theta) = 0$$

which is the equation for minimizing the log-likelihood, so the solution of this equation(if it exists) is the same as the Maximum Likelihood Estimate(MLE) of θ.

If $\alpha \to 1$, the corresponding estimator is called the L_2 divergence estimate. But this estimator is highly robust and we will see why we don't use such a good estimator in practice.

4 Asymptotic Properties

One of the most desirable property for any good estimator is that it is also asymptotically very good, in the sense that it is consistent and have smaller variance relative to other and if possible the order of convergence is bigger i.e. the rate of convergence is much faster than the others. Now our estimator $\hat{\theta}$ is obtained by solving the equation

$$H_n(\hat{\theta}) = \min_{\theta\in\Theta} H_n(\theta)$$

Let us assume that the unique solution of the above equation exists and it is also independent of the index i, say θ^g. Now let us assume a new equation or more mathematically the i^{th} element from the summation format of the objective function i.e.

$$H^{(i)}(\theta) = \int f_i(y,\theta)^{1+\alpha}dy - (1+\frac{1}{\alpha})\int f_i(y,\theta)^{\alpha}g_i(y)dy \quad \forall i = 1,2,....$$

Now the solution θ^g is also a solution of the equation

$$\nabla H^{(i)}(\theta) = 0 \quad i.e. \nabla H^{(i)}(\theta^g) = 0 \quad \forall i = 1,2,....$$

We also define a matrix $J^{(i)}$, $\forall i = 1,2,....$ for each i, of order $p\times p$ (assuming the θ vector is of order $p\times 1$) by,

$$J^{(i)}_{kl} = \frac{1}{1+\alpha}E_i[\nabla_{kl}V_i(y_i,\theta)]$$

where $J^{(i)}_{kl}$ is the $(k,l)^{th}$ element of the matrix $J^{(i)}_{kl}$ and ∇_{kl} denotes the partial derivative corresponding to the indicated component of θ.

We further define the quantities

$$\Psi_n = \frac{1}{n}\sum_{i=1}^{n} J^{(i)}$$

and

$$\Omega_n = \frac{1}{n}\sum_{i=1}^{n} Var_{g_i}[\nabla V_i(y_i,\theta)]$$

$$J^{(i)} = \int u_i\left(y;\theta^g\right) u_i^T\left(y;\theta^g\right) f_i^{1+\alpha}\left(y;\theta^g\right) dy$$
$$- \int \left\{\nabla u_i\left(y;\theta^g\right) + \alpha u_i\left(y;\theta^g\right) u_i^T\left(y;\theta^g\right)\right\} \left\{g_i(y) - f_i\left(y;\theta^g\right)\right\} f_i\left(y;\theta^g\right)^\alpha dy$$

and

$$\Omega_n = \frac{1}{n}\sum_{i=1}^{n}\left[\int u_i\left(y;\theta^g\right) u_i^T\left(y;\theta^g\right) f_i\left(y;\theta^g\right)^{2\alpha} g_i(y)dy - \xi_i\xi_i^T\right]$$

where,

$$\xi_i = \int u_i(y,\theta^g) f_i(y,\theta^g)^\alpha g_i(y)dy$$

Now we will consider the basic assumptions for which the asymptotic properties will follow for our DPD estimator.

- The support $\chi = \{y|f_i(y;\theta) > 0\}$ is independent of i and θ $\forall i$; the true distributions G_i are also supported on χ $\forall i$.
- There is an open subset of ω of the parameter space Θ, containing the best fitting parameter θ^g such that for almost all $y \in \chi$, and all $\theta \in \Theta$, all i = 1, 2, . . ., the density $f_i(y;\theta)$ is thrice differentiable with respect to θ and the third partial derivatives are continuous with respect to θ
- For i = 1,2,..., the integrals $\int f_i(y;\theta)^{1+\alpha}dy$ and $\int f_i(y;\theta)^\alpha gi(y)dy$ can be differentiated thrice with respect to θ, and the derivatives can be taken under the integral sign.
- For each i = 1,2,..., the matrices $J^{(i)}$ are positive definite and
$$\lambda_0 = \inf_n[min\ eigenvalue\ of\ \psi_n] > 0$$
- There exists a function $M_{jkl}^{(i)}(y)$ such that
$$|\nabla_{jkl}V_i(y;\theta)| \le M_{jkl}^{(i)}(y) \quad \forall\theta\in\Theta, \quad \forall i$$
where
$$\frac{1}{n}\sum_{i=1}^{n} E_{g_i}\left[M_{jkl}^{(i)}(Y)\right] = O(1) \quad \forall j,k,l.$$
- For all j,k, we have
$$\lim_{N\to\infty}\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n} E_{g_i}[|\nabla_j V_i(Y;\theta)|I(|\nabla_j V_i(Y;\theta)| > N)] = 0$$
$$\lim_{N\to\infty}\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n} E_{g_i}[|\nabla_{jk}V_i(Y;\theta) - E_{g_i}(\nabla_{jk}V_i(Y;\theta))|$$
$$\times I(|\nabla_{jk}V_i(Y;\theta) - E_{g_i}(\nabla_{jk}V_i(Y;\theta))| > N)] = 0$$
where I(B) denote the indicator random variable of the event B.

- For all $\epsilon > 0$, we have

$$\lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} E_{g_i} \left[\left\| \Omega_n^{-1/2} \nabla V_i(Y;\theta) \right\|^2 I\left(\left\| \Omega_n^{-1/2} \nabla V_i(Y;\theta) \right\| > \epsilon\sqrt{n} \right) \right] \right\} = 0$$

4.1 Theorem

. Under the above assumptions, the following results hold:

i. There exists a consistent sequence θ_n of roots to the minimum DPD estimating equation(5).

ii. The asymptotic distribution of $\Omega^{-\frac{1}{2}} \Psi_n [n(\theta_n - \theta^g)]$ is pdimensional normal with (vector) mean 0 and covariance matrix I_p, the p-dimensional identity matrix.

Now from the second statement of theorem 3.1, we can conclude that

$$\sqrt{n}(\theta_n - \theta^g) \sim N_p(0, \Psi_n^{-1} \Omega \Psi_n^{-1})$$

From this statement, we have a better position to conclude that the order of convergence for the sequence of parameter θ_n is $O(\frac{1}{n})$ which is the same as the rate of convergence of the corresponding MLE estimator(which is of order $O(\frac{1}{\sqrt{n}})$). In the next section, we will consider the robustness of our proposed estimator.

5 Influence Function

In robust analysis, the influence function plays a significant role in finding out the robustness of the estimator. By observing the nature of the influence function, analytically or graphically, we can safely comment on whether the estimator is robust and if it is a robust estimator then how much it is (compared to the other estimators).

Let us define,

$$\sum_{i=1}^{n} H^{(i)}(\theta) = \sum_{i=1}^{n} \left[\int f_i(y;\theta)^{1+\alpha} dy - \frac{1+\alpha}{\alpha} \int f_i(y;\theta)^{\alpha} dG_i(y) \right] \forall i = 1, 2, ...$$

under appropriate differentiability conditions, as the solution of the estimating equation

$$\sum_{i=1}^{n} \nabla H^{(i)}(\theta) = 0,$$

$$\implies (1+\alpha) \sum_{i=1}^{n} \left[\int f_i(y;\theta)^{1+\alpha} u_i(y;\theta) dy - \int f_i(y;\theta)^{\alpha} u_i(y;\theta) g_i(y) dy \right] = 0$$

$$\implies \sum_{i=1}^{n} \left[\int f_i(y;\theta)^{1+\alpha} u_i(y;\theta) dy - \int f_i(y;\theta)^{\alpha} u_i(y;\theta) g_i(y) dy \right] = 0$$

Now we want to analyze the above function through the idea given by Huber. The main idea is the contamination of the densities $g_{i,\epsilon} = (1-\epsilon)g_i + \epsilon \wedge_{t_i}$ where $\wedge_{t_i}$ is the degenerate distribution at the the point of contamination t_i and G_i denotes corresponding distribution function for all $\forall i = 1(1)n$. let $\theta_c^{i_0} = T_\alpha\left(G_1, \ldots, G_{i_0-1}, G_{i_0,\epsilon}, \ldots, G_n\right)$ be the minimum density power divergence functional with contamination only in the i_0-th direction. Now substitute $\theta_\epsilon^{i_0}$ and $g_{i_0,e}$ in place of θ and g_{i_0} respectively in the estimating equation; differentiating with respect to ϵ and evaluating at $\epsilon = 0$, we then get the influence function of the functional which considers contamination only along the i_0-th direction to be where $\xi_i = \int u_i(y;\theta) f_i(y;\theta) g_i(y) dy$. Similarly, letting $\theta_e = T_\alpha\left(G_{1,e}, \ldots, G_{n,c}\right)$ and proceeding similarly, we get the influence function with contamination at all the data-points as

$$IF\left(t_1, \ldots, t_n, T_\alpha, G_1, \ldots, G_n\right) = \Psi_n^{-1} \frac{1}{n} \sum_{i=1}^{n} \left[f_i\left(t_i;\theta\right)^\alpha u_i\left(t_i;\theta\right) - \xi_{i_0}\right].$$

Following these approaches, we will define some influence function-based gross summary measures for our non-homogeneous setup. The simplest one is the (unstandardized) **gross-error sensitivity** of the functional T_α at the true distributions $G_1, \ldots, G_n$ considering contamination only in the i_0^{th} direction, which is defined as

$$\begin{aligned}
\gamma_{i_0}^u\left(T_\alpha, G_1, \ldots, G_n\right) &= \sup_t \left\{\left\|IF_{i_0}\left(t, T_\alpha, G_1, \ldots, G_n\right)\right\|\right\} \\
&= \frac{1}{n} \sup_t \left\{\left[f_{i_0}(t;\theta)^\alpha u_{i_0}(t;\theta) - \xi_{i_0}\right]^T \Psi_n^{-2} \left[f_{i_0}(t;\theta)^\alpha u_{i_0}(t;\theta) - \xi_{i_0}\right]\right\}^{\frac{1}{2}}.
\end{aligned}$$

However it is not invariant to scale transformation of the individual parameter components. Whenever, the asymptotic variance of the corresponding MDPE exists, we can overcome this problem by considering the Self-Standardized Sensitivity. For contamination along the i_0-th direction only, this is defined as

$$\begin{aligned}
\gamma_{i_0}^s(T_\alpha, G_1, \ldots, G_n) &= \sup_t \left\{IF_{i_0}\left(t, T_\alpha, G_1, \ldots, G_n\right)^T \left(\Psi_n^{-1}\Omega_n\Psi_n^{-1}\right)^{-1} IF_{i_0}\left(t, T_\alpha, G_1, \ldots, G_n\right)\right\}^{\frac{1}{2}} \\
&= \frac{1}{n} \sup_t \left\{\left[f_{i_0}(t;\theta)^\alpha u_{i_0}(t;\theta) - \xi_{i_0}\right]^T \Omega_n^{-1} \left[f_{i_0}(t;\theta)^\alpha u_{i_0}(t;\theta) - \xi_{i_0}\right]\right\}^{\frac{1}{2}}
\end{aligned}$$

When we have contamination in all the directions, we can define the (unstandardized) gross-error sensitivity $\gamma^u\left(T_\alpha, G_1, \ldots, G_n\right)$ and the self-standardized sensitivity $\gamma^s\left(T_\alpha, G_1, \ldots, G_n\right)$ using equation (4.6) and (4.8) respectively with $IF_{i_0}\left(t, T_\alpha, G_1, \ldots, G_n\right)$ replaced by $IF(t_1, ..., t_n, T_\alpha, G_1, \ldots, G_n$ and taking supremum over all possible $t_1, \ldots, t_n$.

6 Breakdown Point of the Location Parameter in the Location-Scale type Model

In this section we will derive the breakdown point of the minimum density power divergence estimator in above set-up of non-homogeneous observations for the location parameter in a special class of models. We will consider the above set-up and assume that

$$f_i(y;\theta) \in \mathcal{F}_{i,\theta} = \left\{\frac{1}{\sigma} f\left(\frac{y - l_i(\mu)}{\sigma}\right) : \theta = (\mu, \sigma) \in \Theta\right\}$$

where $l_i(\cdot)$ is some one-to-one function for each $i = 1, \ldots, n$. We consider the breakdown point of the estimator of the location parameter μ at the model and will assume the scale-parameter σ to be fixed (for example, σ can be substituted with any suitable robust scale estimator) and the true data generating densities g_i to belongs to the model family $\mathcal{F}_{i,\theta}$; thus, for each $i, g_i(y) = \frac{1}{\sigma} f\left(\frac{y - l_i(\mu_a)}{\sigma}\right)$, where μ_g is the true value of the location parameter μ. For given σ, the minimum density power divergence estimator of μ is defined as

$$T_\alpha^\mu\left(G_1, \ldots, G_n\right) = \arg\min_\mu \frac{1}{n} \sum_{i=1}^{n} d_\alpha\left(g_i(\cdot), f_i(\cdot; \theta)\right).$$

Assume n to be fixed and consider the contamination models

$$H_{i,\epsilon,m} = (1 - \epsilon) G_i + \epsilon K_{i,m},$$

for each i where $\{K_{i,m}\}$ is a sequence of contaminating distributions. Denote the corresponding densities by $h_{i,\epsilon,m}, g_i$ and $k_{i,m}$. Following Simpson (1987) , we say that there is a breakdown in T_α^μ for ϵ level contamination if there exists sequences $K_{i,m}$ such that

$$\left|T_\alpha^\mu\left(H_{1,\epsilon,m}, \ldots, H_{n,\epsilon,m}\right) - T_\alpha^\mu\left(G_1, \ldots, G_n\right)\right| \to \infty \quad \text{as} \quad m \to \infty.$$

Here we will use a generalization of the argument used by Park and Basu (2004) [11] to derive the breakdown of the minimum disparity estimators. Recall that we can also write the density power divergence in equation (2.1) as

$$d_\alpha(g, f) = \int f^{1+\alpha} C_\alpha(g/f) = \int f^{1+\alpha} C_\alpha(\delta + 1)$$

where $\delta = g/f - 1$ and

$$C_\alpha(\delta + 1) = \frac{1}{\alpha}\left[\alpha - (1 + \alpha)(\delta + 1) + (\delta + 1)^{1+\alpha}\right].$$

In the above integral we have suppressed the dummy variable and the differential for simplicity of notation. Note that $C_\alpha(0) = 1$. Define $D_\alpha(g, f) = f^{1+\alpha} C_\alpha(g/f)$. Whenever $\alpha > 0$, we have

$$D_\alpha(g, 0) = \lim_{f \to 0} D_\alpha(g, f) = \lim_{f \to 0}\left[f^{1+\alpha} - \frac{1+\alpha}{\alpha} f^\alpha g + \frac{1}{\alpha} g^{1+\alpha}\right] = \frac{1}{\alpha} g^{1+\alpha}.$$

We also utilize useful results based on the special structure of the location-scale type model considered here. For example, note that

$$\int \left\{\frac{1}{\sigma} f\left(\frac{y - l_i(\mu)}{\sigma}\right)\right\}^{1+\alpha} dy = \frac{1}{\sigma^\alpha} \int \{f(x)\}^{1+\alpha} dx = \frac{1}{\sigma^\alpha} M_f^\alpha, \quad \text{say}$$

which is independent of the location parameter μ and the index i. In addition, we have the crucial lemma given below.

Lemma 5.1. Assume that $\alpha > 0$ and fix any i. Then for any two densities g_i, h_i in the location-scale model $\mathcal{F}_{i,\theta}$ with fixed $\sigma > 0$ and any $\epsilon \in (0,1)$, the integral $\int D_\alpha(\epsilon g_i, h_i)$ is minimized when $g_i = h_i$.

We are now in a position to state and prove our main result on breakdown. First we provide the necessary set of assumptions.

(BP1) For each $i = 1, \ldots, n$, $\int \min\{f_i(y;(\mu,\sigma)), k_{i,m}(y)\} \to 0$ as $m \to \infty$ uniformly for $|\mu| \leq c$ for any fixed c. That is, the contamination distribution is asymptotically singular to the true distribution and to specified models within the parametric family.

(BP2) For each $i = 1, \ldots, n$, $\int \min\{f_i(y;(\mu_g,\sigma)), f_i(y;(\mu_m,\sigma))\} \to 0$ as $m \to \infty$ if $|\mu_m| \to \infty$ as $m \to \infty$. That is, large values of the parameter μ give distributions that become asymptotically singular to the true distribution.

(BP3) Let $C_\alpha(\cdot)$ be as . For each $i = 1, \ldots, n$, the contaminating sequence $\{k_{i,m}\}$ is such that

$$d_\alpha(\epsilon k_{i,m}(.), f_i(.;\theta)) \geq d_\alpha(\epsilon f_i(.;\theta), f_i(.;\theta)) = \frac{C_\alpha(\epsilon)}{\sigma^\alpha} M_f^\alpha$$

for any $\theta \in \Theta$ and $0 < \epsilon < 1$ and

$$\limsup_{m\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int k_{i,m}^{1+\alpha} \leq \frac{M_f^\alpha}{\sigma^\alpha}.$$

We then have the following theorem. Theorem 5.2. Assume that $\alpha > 0$. Then under the assumptions (BP1)-(BP3) above, the asymptotic breakdown point ϵ^* of the minimum DPD functional T_α^μ of the location parameter μ is at least $\frac{1}{2}$ at the location-scale set up of (5.1) for fixed scale parameters.

The above theorem establishes that the minimum DPD procedure generates estimators with high breakdown points for all $\alpha > 0$. For the i.i.d. setup obtained by letting $f_i(y;\theta) = f_\theta(y)$, Theorem 5.2 directly yields the following Corollary.

Suppose independent and identically distribution data are obtained from the location scale model $\mathcal{F}_\theta = \left\{\frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right) : \theta = (\mu,\sigma) \in \Theta\right\}$ and assume that the scale parameter σ is fixed. Then under assumptions (BP1)(BP3) and for all $\alpha > 0$, the minimum density power divergence estimator of the location parameter μ has an asymptotic breakdown point of at least $\frac{1}{2}$ at the model.

The above result may be contrasted with the Basu et al. (1998 [2], Section 4.3) result which gives the simultaneous location and scale breakdown point of the minimum DPD estimator to be $\alpha/(1+\alpha)^{3/2}$. The remarks give us some justification for assumptions (BP1)-(BP3).

Suppose that the contaminating densities $\{k_{i,m}\}$ belongs to the model presented in equation (5.1), and satisfies the set up of this section for all i and all m. Then the following results are seen to be true. 1 Assumption (BP3) holds. The second part of the assumption holds trivially and the first part of the assumption holds by Lemma 5.1. 2 Let $k_{i,m} = f_i(y;(\mu_m,\sigma))$ and suppose that $|\mu_m| \to \infty$ as $m \to \infty$. Then assumption (BP2) implies assumption (BP1). 3 If we assume that $f(\cdot) = \phi(\cdot)$ in the model represented by equation (5.1) where $\phi(\cdot)$ is the univariate normal density, assumption (BP2) also holds trivially.

We expect that it will be possible to prove the breakdown result in Theorem 5.2 under conditions where a weaker version of (BP3) will suffice but we do not have a proof at this point.

7 Normal Linear Regression

As we have defined earlier that we will consider the case for the distribution where the all the data coming form the different distributions and the they are interrelated by a single common unknown parameter(may be vector). By seeing this the most obvious thing that may come in our mind is the linear regression. Actually we can imagine the theory of the multiple linear regression.

In linear regression the proposed model looks like,

$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + ... + \beta_p * x_{ip} + \epsilon_i$$

where $\beta_1, \beta_2, ..., \beta_p$ are p unknown parameters and $x_{i1}, x_{i2}, ..., x_{ip}$ are i-th component of p regressors and y_i be the i-th component of the response variable and ϵ_i be the i-th random component of the $\forall i = 1(1)n$.

Now as we compare the theory of this method to the theory of MLE,we must have to make some distributional assumption on the random component. So, we will assume the most basic assumption i.e. the errors follow $N(0, \sigma^2)$ distribution independently and identically. This will imply that the distribution y_i is $N(\boldsymbol{x_i'\beta}, \sigma^2)$ where $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ and $\boldsymbol{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})$. So there is a single common parameter vector named $(\boldsymbol{\beta}, \sigma^2)$. Usually in Mle we actually get the same estimate of $\boldsymbol{\beta}$'s as from the least square optimization. We will not going into the deep theory of least square optimization, rather than we will implement the corresponding theory from the general theory of minimum Density power Divergence Estimator. Let us first define the i-th component from the summation of the objective function,

$\frac{1}{n}\sum_{i=1}^{n} V_i(y_i; \theta, x_i) = \frac{1}{(2\pi)^{\alpha/2}\sigma^{\alpha}\sqrt{1+\alpha}} \quad - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^{\alpha}}\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha\left(y_i - x_i^T\beta\right)^2/(2\sigma^2)}$ Thus, our objective function to be minimized becomes

$$\frac{1}{n}\sum_{i=1}^{n} V_i\left(y_i; \theta, x_i\right) = \frac{1}{(2\pi)^{\alpha/2}\sigma^{\alpha}\sqrt{1+\alpha}} - \frac{1+\alpha}{\alpha}\frac{1}{(2\pi)^{\alpha/2}\sigma^{\alpha}}\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha\left(y_i - x_i^T\beta\right)^2/\left(2\sigma^2\right)},$$

which is exactly the same as the one suggested by Basu et al. (1998) [2] for linear regression and the equation considered by Durio and Isaia (2011 [4], equation (3)). Letting $\nabla_j, j = 1, \ldots, p$ represent the partial derivative with respect to β_j we get

$$\nabla_j V_i\left(y_i; \theta, x_i\right) = -\frac{1+\alpha}{(2\pi)^{\alpha/2}\sigma^{\alpha+2}} e^{-\frac{\alpha\left(y_i - x_i^T\beta\right)^2}{2\sigma^2}} \left(y_i - x_i^T\beta\right) x_{ij} \quad \forall j = 1, \ldots, p$$

and the partial derivative with respect to σ^2 is then

$$\begin{aligned}
&\nabla_{p+1} V_i\left(y_i;\theta,x_i\right)\\
&=-\frac{1}{(2\pi)^{\alpha/2}}\left[\frac{\alpha}{2\sigma^{\alpha+2}\sqrt{1+\alpha}}-\frac{(1+\alpha)}{2\sigma^{\alpha+2}}e^{-\frac{\alpha\left(y_i-x_i^T\beta\right)^2}{2\sigma^2}}\left\{1-\frac{\left(y_i-x_i^T\beta\right)^2}{\sigma^2}\right\}\right].
\end{aligned}$$

Thus we get the estimating equation to be

$$\begin{aligned}
&\sum_{i=1}^{n} x_{ij}\left(y_i-x_i^T\beta\right)e^{-\frac{\alpha\left(y_i-x_i^T\beta\right)^2}{2\sigma^2}}=0 \quad \forall j=1,\ldots,p\\
&\sum_{i=1}^{n}\left[1-\frac{\left(y_i-x_i^T\beta\right)^2}{\sigma^2}\right]e^{-\frac{\alpha\left(y_i-x_i^T\beta\right)^2}{2\sigma^2}}=\frac{\alpha}{(1+\alpha)^{\frac{3}{2}}}.
\end{aligned}$$

We can then solve these estimating equations numerically to obtain the estimates of θ. Let us denote these estimators by $\hat{\theta}^T=\left(\hat{\beta}^T,\hat{\sigma}^2\right)$.

7.1 Asymptotic Efficiency

As we have discussed the idea of the asymptotic statistics earlier, here we also calculate different expressions of J, Ω_n,Θ_n. But here our calculation may be easier just because of our simplified assumption of the random error or equivalently the non-homogeneous but independent assumption of the response variable.
Let,

$$u_i\left(y_i;\theta\right)=\begin{pmatrix}\frac{\left(y_i-x_i^T\beta\right)}{\sigma^2}x_i\\ \frac{\left(y_i-x_i^T\beta\right)^2}{2\sigma^4}-\frac{1}{2\sigma^2}\end{pmatrix}.$$

Thus a routine calculation shows that the matrix $J^{(i)}$ is given by

$$J^{(i)}=\int u_i(y;\theta)u_i(y;\theta)^T f_i(y;\theta)^{1+\alpha}dy=\begin{pmatrix}\zeta_\alpha x_i x_i^T & 0\\ 0 & \varsigma_\alpha\end{pmatrix}$$

where

$$\begin{aligned}
\zeta_\alpha&=(2\pi)^{-\frac{\alpha}{2}}\sigma^{-(\alpha+2)}(1+\alpha)^{-\frac{3}{2}}\\
\varsigma_\alpha&=(2\pi)^{-\frac{\alpha}{2}}\sigma^{-(\alpha+4)}\frac{1}{4}\left(\frac{2+\alpha^2}{(1+\alpha)^{\frac{5}{2}}}\right).
\end{aligned}$$

Therefore, we have a simplified form for the matrix Ψ_n as

$$\Psi_n=\begin{pmatrix}\frac{\zeta_\alpha}{n}\left(X^TX\right) & 0\\ 0 & \varsigma_\alpha\end{pmatrix}$$

where $X^T=(x_1,\ldots,x_n)_{p\times n}$. Similarly, we get

$$\xi_i=\int u_i(y;\theta)f_i(y;\theta)^{1+\alpha}dy=\begin{pmatrix}0\\ -\frac{\alpha}{2}\zeta_\alpha\end{pmatrix}$$

and hence

$$\Omega_n = \begin{pmatrix} \frac{\zeta_{2\alpha}}{n}\left(X^T X\right) & 0 \\ 0 & \varsigma_{2\alpha} - \frac{\alpha^2}{4}\zeta_\alpha^2 \end{pmatrix}$$

Using the above, we are now in a position to derive the asymptotic distributions of the minimum DPD estimator of the regression coefficients and error variances under the assumptions or the regularity conditions in the theory of asymptotic efficiency. We first present some mild conditions on the given values of the independent variables, under which these assumptions may be shown to hold.

(R1) The values of x_i 's are such that for all j, k, and l

$$\sup_{n>1} \max_{1\le i\le n} |x_{ij}| = O(1), \quad \sup_{n>1} \max_{1\le i\le n} |x_{ij}x_{ik}| = O(1),$$

and

$$\frac{1}{n}\sum_{i=1}^{n} |x_{ij}x_{ik}x_{il}| = O(1).$$

(R2) The matrix $X^T = (x_1, \ldots, x_n)_{p\times n}$ satisfies

$$\inf_n \left[\text{min eigenvalue of } \frac{\left(X^T X\right)}{n}\right] > 0,$$

which also implies that the matrix X has full column rank, and

$$n \max_{1\le i\le n} \left[x_i^T \left(X^T X\right)^{-1} x_i\right] = O(1).$$

Then the following lemma is easily seen to be true. Lemma 6.1. Consider the set-up of the normal linear regression model and assume that the true data-generating density belongs to the model family. Then the conditions $(R1)$ and $(R2)$ imply assumptions $(A1) - (A7)$.

Note that the conditions (R1) and (R2) on the x_i 's mainly says that their values remain bounded in large samples and the spectrum of the corresponding sum-product matrix $\left(X^T X\right)$ remains bounded away from zero. With these conditions, the asymptotic distribution of the minimum density power divergence estimators of the parameters of the linear regression model are derived in the following theorem:

Theorem 6.2. Under the set-up of the normal linear regression model considered here, assume that the true data generating density belongs to the model family and the given values of the independent variables satisfies assumptions (R1) and (R2). Then, (i) There exists a consistent sequence as $\hat{\theta}^T = \left(\hat{\beta}^T, \hat{\sigma}^2\right)$ of roots to the minimum DPD estimating equations (6.3) and (6.4). (ii) The asymptotic distributions of $\hat{\beta}$ and $\hat{\sigma}^2$ are independent. (iii) The asymptotic distribution of $\left(X^T X\right)^{\frac{1}{2}}(\hat{\beta} - \beta)$ is a p-dimensional normal with a mean (vector) and covariance matrix $v_\alpha^\beta I_p$ and $\sqrt{n}\,(\hat{\sigma}^2 - \sigma^2)$ follows a normal distribution with mean 0 and variance v_α^e, where

$$\mathrm{v}_\alpha^\beta = \frac{\zeta_{2\alpha}}{\zeta_\alpha^2} = \sigma^2\left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{\frac{s}{2}}$$

$\mathrm{v}_\alpha^e = \frac{\varsigma_{2\alpha} - \frac{\alpha^2}{4}\zeta_\alpha^2}{\varsigma_\alpha^2} = \frac{4\sigma^4}{(2+\alpha^2)^2}\left[2\left(1+2\alpha^2\right)\left(1+\frac{\alpha^2}{1+2\alpha}\right)^{\frac{1}{2}} - \alpha^2(1+\alpha)^2\right].$

Note that by substituting $\alpha = 0$ in the expression of the asymptotic variances of the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ in the above theorem, we will get precisely the same results as obtained for their maximum likelihood estimates.

We will now look at the Asymptotic Relative Efficiency (ARE) of the minimum density power divergence estimators concerning the (fully efficient) maximum likelihood estimator. The ARE of the estimator $\hat{\beta}$ of the regression coefficient $\beta = (\beta_1, \ldots, \beta_p)$ is the same for all the β_i 's and is given by

$$\frac{v_0^\beta}{v_\alpha^\beta} \times 100 = \left(1+\frac{\alpha^2}{1+2\alpha}\right)^{-\frac{2}{2}} \times 100$$

Similarly, the asymptotic relative efficiency of the estimator $\hat{\sigma}^2$ of the error variance is given by

$$\frac{v_0^e}{v_\alpha^e} \times 100 = \frac{(2+\alpha^2)^2}{2}\left[2\left(1+2\alpha^2\right)\left(1+\frac{\alpha^2}{1+2\alpha}\right)^{\frac{\pi}{2}} - \alpha^2(1+\alpha)^2\right]^{-1} \times 100.$$

Table 1 presents the asymptotic relative efficiencies of these estimators for various values of α. From the table, it is easy to see that the loss of efficiency is quite small for small values of α. It is also interesting to note that the ARE of

Table 1: The Asymptotic Relative Efficiencies of the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ for various values of the tuning parameter α

α	0	0.01	0.02	0.05	0.10	0.15	0.25	0.50	0.75	1.00
$ARE(\hat{\beta})$	100	99.99	99.94	99.66	98.76	97.46	94.06	83.81	73.76	64.95
$ARE\left(\hat{\sigma}^2\right)$	100	99.97	99.88	99.32	97.56	95.05	88.84	73.06	61.53	54.11

the minimum DPD estimator of the regression coefficient β is the same as the ARE of the minimum DPD estimator of the normal mean parameter and ARE of the minimum DPD estimator of error variance σ^2 is the same as the ARE of the normal variance are reported in Basu et al. (1998) .

7.2 Influence function and sensitivities

Now we will define or more mathematically, will simplify the general theory of Influence function and sensitivities of the non-homogeneous but independent case to the multiple linear regression setup. Note that from the expression of Ψ_n(previously defined), we get

$$\Psi_n^{-1} = \begin{pmatrix} \frac{n}{\zeta_\alpha}\left(X^T X\right)^{-1} & 0 \\ 0 & \frac{1}{\zeta_\alpha} \end{pmatrix}$$

Then using the expression of u_i and $\xi_i$respectively, we get the influence function of the estimator T_α with contamination at the direction i_0 which is given by

$$IF_{i_0}\left(t_{i_0}, T_\alpha, G_1, \ldots, G_n\right) = \begin{pmatrix} \frac{1}{\zeta_\alpha}\left(X^TX\right)^{-1}\frac{\left(t_{i_0}-x_{i_0}^T\beta\right)}{\sigma^2}x_{i_0}f_{i_0}\left(t_{i_0};\theta\right)^\alpha \\ \frac{1}{n\varsigma_\alpha}\left[\left\{\frac{\left(t_{i_0}-x_{i_0}^T\beta\right)^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right\}f_{i_0}\left(t_{i_0};\theta\right)^\alpha+\frac{\alpha}{2}\zeta_\alpha\right] \end{pmatrix}.$$

Simplifying, the influence function for the estimator T_α^β of the regression coefficients with contamination only in i_0-th data-point only becomes

$$IF_{i_0}\left(t_{i_0}, T_\alpha^\beta, G_1, \ldots, G_n\right) = (1+\alpha)^{\frac{3}{2}}\left(X^TX\right)^{-1}x_{i_0}\left(t_{i_0}-x_{i_0}^T\beta\right)e^{-\frac{\alpha\left(t_{i_0}-x_{i_0}^T\beta\right)^2}{2\sigma^2}}$$

and the influence function for the estimator T_α^σ of the error variance with contamination in the i_0-th data-point only becomes

$$\begin{aligned} IF_{i_0}\left(t_{i_0}, T_\alpha^\sigma, G_1, \ldots, G_n\right) = &\frac{2(1+\alpha)^{\frac{5}{2}}}{n\left(2+\alpha^2\right)}\left\{\left(t_{i_0}-x_{i_0}^T\beta\right)^2-\sigma^2\right\}e^{-\frac{\alpha\left(t_{i_0}-x_{i_0}^T\beta\right)^2}{2\sigma^2}} \\ &+\frac{2\alpha(1+\alpha)^2}{n\left(2+\alpha^2\right)} \end{aligned}$$

Since the functions se^{-s^2} and $s^2e^{-s^2}$ are both bounded in $s \in R$, both the influence functions in $IF_{i_0}\left(t_{i_0}, T_\alpha^\sigma, G_1, \ldots, G_n\right)$ and $IF_{i_0}\left(t_{i_0}, T_\alpha^\beta, G_1, \ldots, G_n\right)$ are bounded in t_{i_0} for all $\alpha > 0$ and for any i_0. This implies that the minimum density power divergence estimators with $\alpha > 0$ will be robust with respect to the outliers in any data-point. However the influence functions are clearly unbounded for $\alpha = 0$ which corresponds to the non-robust maximum likelihood estimators.

Similarly the influence function for the estimator T_α^β of the regression coefficients with contamination in all data-point can be shown to be

$$IF\left(t_1, \ldots, t_n, T_\alpha^\beta, G_1, \ldots, G_n\right) = (1+\alpha)^{\frac{3}{2}}\left(X^TX\right)^{-1}\sum_{i=1}^n x_i\left(t_i-x_i^T\beta\right)e^{-\frac{a\left(t_i-\varepsilon_i^T\beta\right)^2}{2\sigma^2}}$$

and the influence function for the estimator T_α^σ of the error variance with contamination in all data-point will be

$$\begin{aligned} IF\left(t_1, \cdot, t_n, T_\alpha^\sigma, G_1, \cdot, G_n\right) = &\frac{2(1+\alpha)^{\frac{5}{2}}}{n\left(2+\alpha^2\right)}\sum_{i=1}^n\left\{\left(t_i-x_i^T\beta\right)^2-\sigma^2\right\}e^{-\frac{\alpha\left(t_i-x_1^T\beta\right)^2}{2\sigma^2}} \\ &+\frac{2\alpha(1+\alpha)^2}{\left(2+\alpha^2\right)}. \end{aligned}$$

Here also the influence functions (6.17) and (6.18) are bounded in t_i 's for all $\alpha > 0$ and unbounded for $\alpha = 0$.

Now let us derive the sensitivities of the estimator of the regression coefficient β to explore the extent of robustness of the estimator with respect to the value of α. Using the form given

in Section 4, the gross-error sensitivity of the estimator T_α^β of β in the case of contamination only in i_0^{th} direction can be found to be

$$\begin{aligned}\gamma_{i_0}^u\left(T_\alpha^\beta, G_1, \ldots, G_n\right) &= \frac{(1+\alpha)^{\frac{3}{2}}}{\sqrt{\alpha}}\sigma e^{-\frac{1}{2}}\left\|\left(X^TX\right)^{-1}x_{i_0}\right\| \quad \text{if } \alpha > 0\\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \alpha = 0\end{aligned}$$

And the self-standardized sensitivity of the estimator T_α^β in the case of contamination only in i_0^{th} direction is given by

$$\begin{aligned}\gamma_{i_0}^s\left(T_\alpha^\beta, G_1, \ldots, G_n\right) &= \frac{(1+\alpha)^{\frac{3}{2}}}{\sqrt{\alpha v_\alpha^\beta}}\sigma e^{-\frac{1}{2}}\sqrt{x_{i_0}^T\left(X^TX\right)^{-1}x_{i_0}} \quad \text{if } \alpha > 0\\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \alpha = 0\end{aligned}$$

or,

$$\begin{aligned}\gamma_{i_0}^s\left(T_\alpha^\beta, G_1, \ldots, G_n\right) &= \frac{(1+2\alpha)^{\frac{3}{4}}}{\sqrt{\alpha}}e^{-\frac{1}{2}}\sqrt{x_{i_0}^T\left(X^TX\right)^{-1}x_{i_0}} \quad \text{if } \alpha > 0\\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \alpha = 0\end{aligned}$$

It is easy to see that both the sensitivities $\gamma_{i_0}^u$ and $\gamma_{i_0}^s$ are decreasing functions of $\alpha > 0$ for any given x_i 's. This implies that in the presence of the outliers in only one direction the robustness of the estimator T_α^β increases as α increases. However, we have seen in Section 6.1 that the asymptotic relative efficiency of the estimator T_α^β of the regression coefficient β decreases as the tuning parameter α increases. Thus the parameter α gives a trade-off between the efficiency and the robustness of the estimator of regression coefficients.

It is interesting to note that besides the tuning parameter α, the sensitivities also depend on the values of the explanatory variable x_i s. Thus the robustness of the estimator T_α^β also depends on the values of x_i 's. Moreover, from the expression of sensitivities, whenever the value of a x_{i_0} becomes far from the center of the data cloud, the value of both the gross-error sensitivity and self-standardized sensitivity increases implying that the robustness decreases. This fact is quite intuitive from the basic concept of outliers in the explanatory variable.

7.3 Breakdown point of the estimator of the regression coefficient

Now we consider the breakdown point of the estimator of the regression coefficients β using the theory developed in Section 5. Note that the regression set-up exactly matches with the set-up considered in Section 5 with $f(\cdot) = \phi(\cdot)$, the standard normal density, $\mu = \beta$, and $l_i(\mu) = l_i(\beta) = x_i^T\beta$ for all i. Since here we are mainly interested in the breakdown of the estimator of β, as, in Section 5 , we will assume that the error variance σ^2 is to be fixed. In practice, we may replace it with any robust estimator that may be assumed to be the same as the true value of σ^2 asymptotically.

Further, it was proved in the previous theorem that the minimum DPD estimator of β is regression equivariant. And it follows from Rousseeuw and Leroy (1987 [12], Theorem 4, page 125) that the finite sample breakdown point of any regression equivariant estimator of β is at most

$$\frac{[(n-p)/2]+1}{n}$$

at all sample Z of size n. Hence the asymptotic breakdown point of β can be at most $\frac{1}{2}$. Thus we get the following theorem giving the maximum asymptotic breakdown of the minimum density power divergence estimator of the regression coefficient β at the model.

Theorem 6.4. Assume the contaminating densities are such that (BP1) and (BP3) hold. Then for any $\alpha > 0$, the asymptotic breakdown point of the minimum density power divergence estimator of the regression coefficient β is exactly $\frac{1}{2}$ at the model.

Also if we assume that the contamination densities also belong to the model family, i.e., $k_{i,m}$ is the $N\left(x_i^T\beta_m, \sigma^2\right)$ density with $|\beta_m| \to \infty$, then by Remark 5.1 the assumptions (BP1) and (BP3) again hold and the above breakdown result follows.

8 Simulation

Now it's time to implement the theoretical results practically, first by the simulation. For this, we need a simulated data-set corresponding to x and y. Let,

8.1 2D Model

```
set.seed(43639)
x = rnorm(1000,0,1)
z = rnorm(10,-10,1)
y = 5 + 0.5*x
x[991:1000] = z
D = data.frame(x,y,class)
class = c(numeric(990)+1,numeric(10)+3)
plot(D$x,D$y,col = D$class,pch = D$class+16,xlab = "x",ylab = "y",
     main = paste("Sctter plot of x and y"))
```

As in the code, We have simulated 1000 data points from the standard normal distribution and named it x. We transpose the data to 5+0.5*x and named it y. Finally, we add 10 data points from N(-10,1) to x and this is our final x. Let us plot the scatter plot of x and y on a 2D axis.

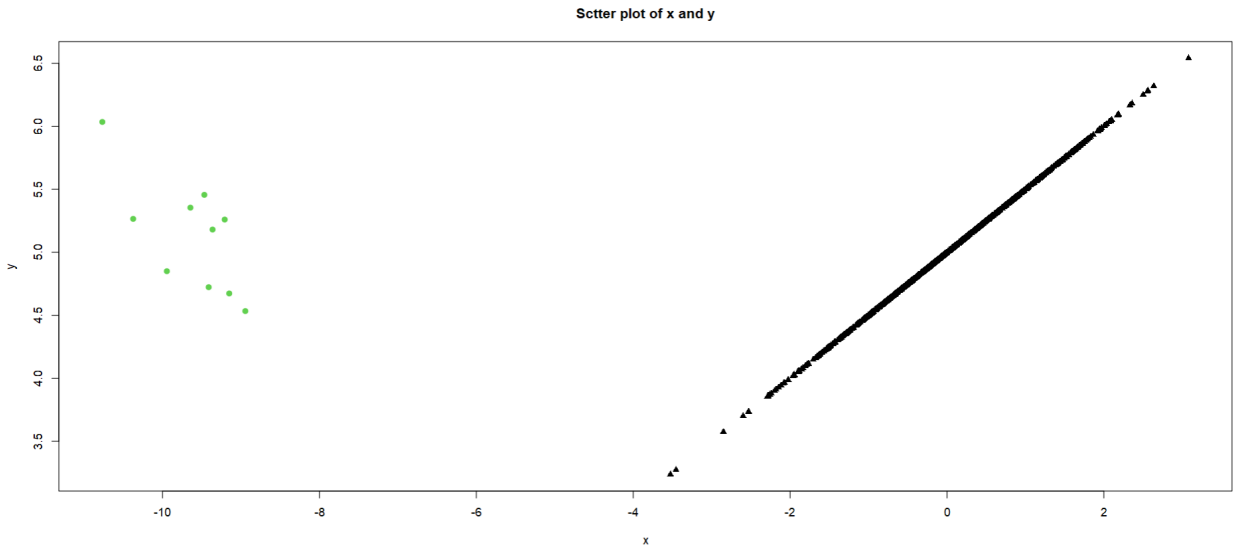

Figure 1: Scatter plot between x and y

As we have seen in the above plot that all the black or original data points are really on a straight line but only 10 green points behave like outliers. Next, we will plot the scatter plot together with the least square regression line.

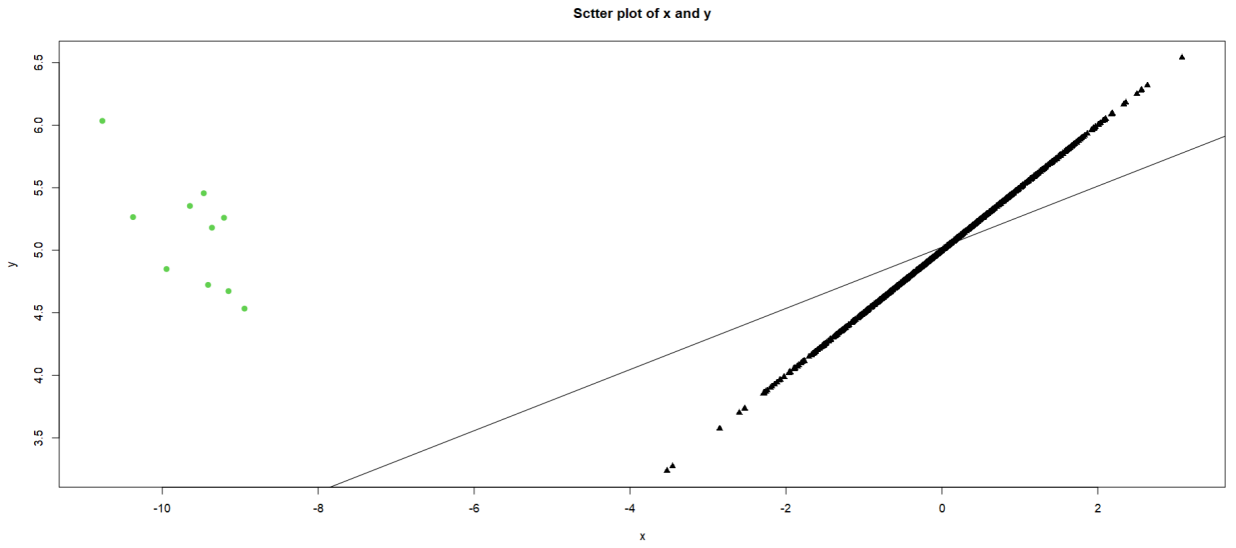

Figure 2: scatter plot with least square regression line

Here it is obvious that the least square regression line is too much dependent on the extreme points or the outliers. Now we will plot the regression line corresponding to our MDPD estimator with $\alpha = 0.05$.

Here we see that the line corresponding to $\alpha = 0.05$ fits the data very well and it is practically not disturbed by the outliers. Now for $\alpha = 0.05$ corresponding efficiency would be more than 99% for both β and σ.

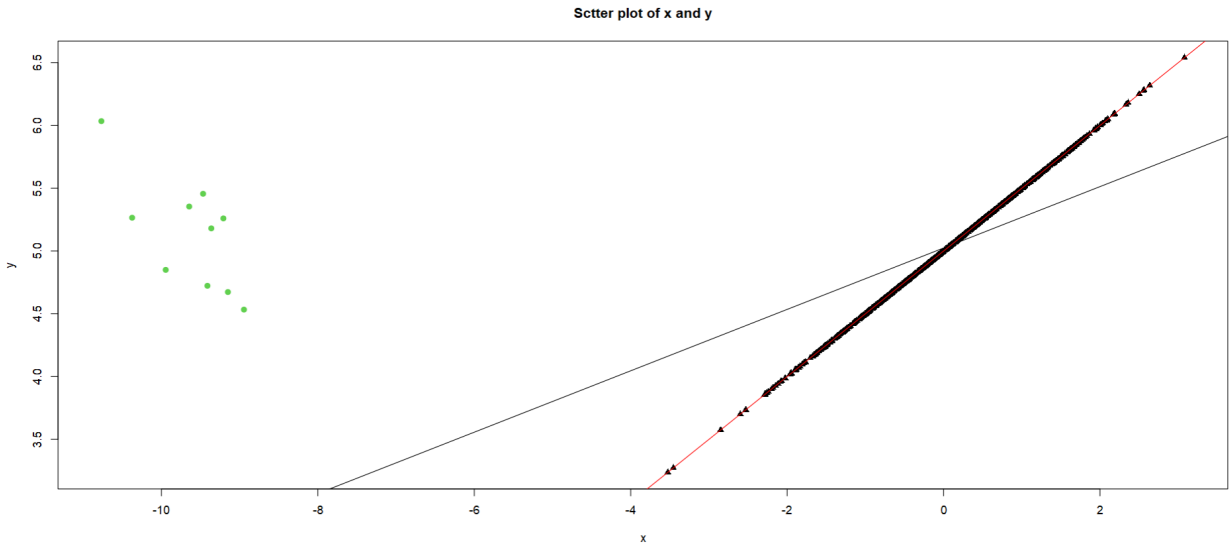

Figure 3: scatter plot with lDPD regression line

8.2 3D Model

Here we build the 3D model to see the effect of the MDPD estimator in a more sophisticated manner.

```
set.seed(65688)
library(rgl)
library(magick)
library(webshot2)
library(LaplacesDemon)
library(mvtnorm)
xm = rmvnorm(1000,c(0,0),diag(2)*10)
zm = rmvnorm(10,c(-10,-50),diag(2))

ym = 5 + (xm[,1]*5)+(xm[,2]*10) + rnorm(1000,-10,10)
xm[991:1000,] = zm
class = c(rep("yellow",990),rep("green",10))
Dm = data.frame(xm,ym,class)


plot3d(Dm,col = Dm$class,size = 5,radius = 2,type = "s")

Modelm = lm(ym~X1+X2,data = Dm)
summary(Modelm)
P = Modelm$coefficients
Data = data.frame(xm,Modelm$fitted.values)
planes3d(a = P[2],b = P[3],c = -1,d = P[1],col = "red",alpha = 0.9)
```

As in the above code, we generate 1000 observations from $N(\mathbf{0}, 10 * I_2)$ and named it xm.

Then we calculate 5 + first column(xm)*5 + second column(xm)*10 and named it ym. Now to create some sort of randomness we add a random component i.e. one random observation from N(-10,10) and add the corresponding component of ym. Again we create some disturbance just created by the outliers from a bi-variate normal distribution with mean(-10,-50) and covariance matrix I_2.

So the three-dimensional scatter plot for this problem is of the form -

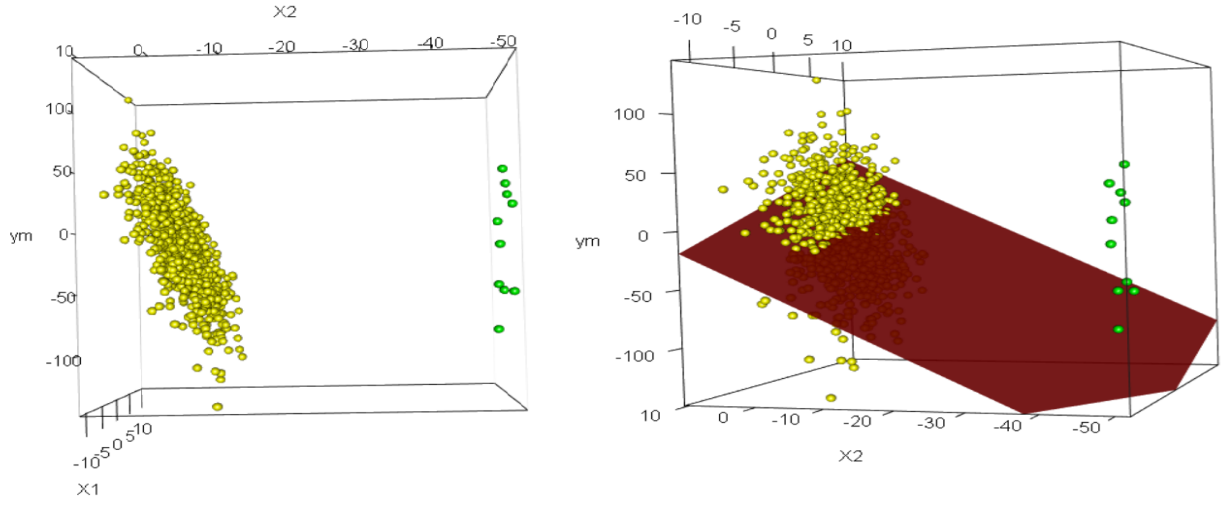

Figure 4: (a):Scatter plot (b):Scatter plot with least square regression plane

Here also only 10 outliers are present in our data marked as green but the least square regression plane is heavily affected by the outliers. So let us define our MDPD estimator for this particular case.

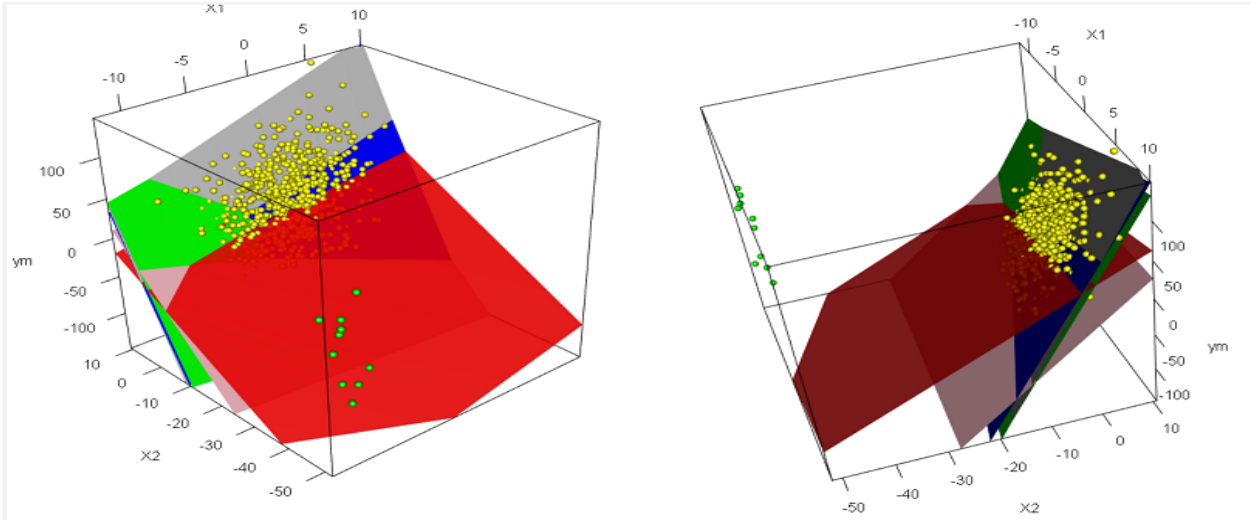

Figure 5: DPD estimators from two different view point

Now we see that the planes are not too easy to visualize but for $\alpha = 0.05$ we get a sufficiently good estimator compared to the least square estimator.

9 Real Data Analysis

As we have seen that the stimulative data examples are parallel to the theory of our analytical results. Now we will see the behavior of these MDPD estimators when the data is a real data example.

9.1 Star Data Example

As our first example, we consider the data for the Hertzsprung-Russell diagram of the star cluster CYG OB1 containing 47 stars in the direction of Cygnus. For this data, the independent variable x is the logarithm of the effective temperature at the surface of the star (Te), and the dependent variable y is the logarithm of its light intensity (L/L_0). As we have from

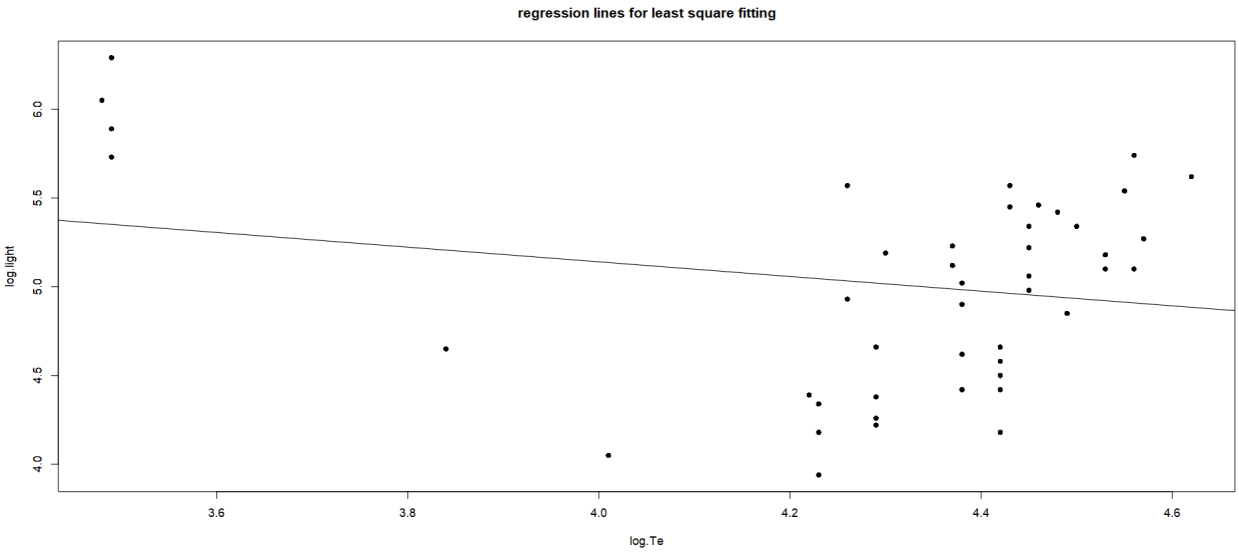

Figure 6: scatter plot with least square regression line

the plot the four stars have very less temperature but they emit too much light. So they do not follow any regular point and they behave like outliers.

As the picture shows that the least square regression line is heavily affected by the outliers and fits the data too worse making the estimation worthless. After that, we will follow the theory of the MDPD estimator to verify our analytical justification graphically.

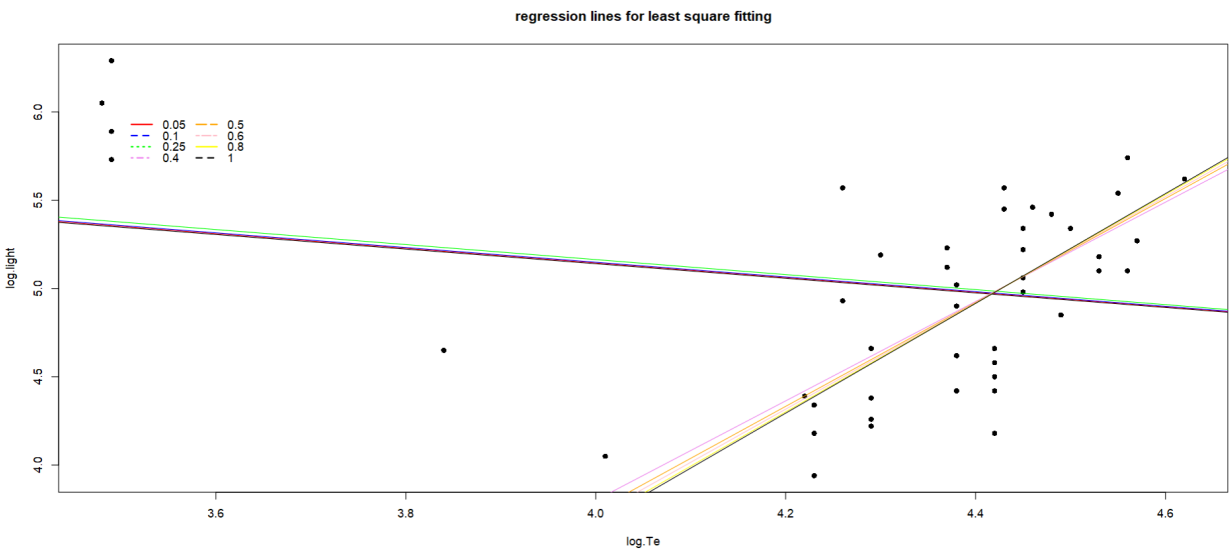

Figure 7: scatter plot with least square and DPD regression line

In this case, the breakdown point is

$$\frac{[n-p]/2}{n} = \frac{[47-2]/2}{47} = 0.48$$

So after that, all the estimates are highly robust and before that, all the data points are highly efficient. So we can choose any value of $\alpha > 0.48$. So any estimate greater than the above cutoff value is highly preferable from a robust point of view. Finally, we will show the estimate of the parameters in its numerical figure.

	0	0.05	0.1	0.25	0.4	0.5	0.6	0.8	1
β_1	6.79	6.80	6.81	6.862	-7.45	-8.03	-8.32	-8.61	-8.76
β_2	-0.41	-0.41	-0.41	-0.42	2.81	2.94	3.01	3.07	3.11
σ	0.56	0.56	0.56	0.58	0.39	0.39	0.39	0.40	0.41

Table 2: Different values of β_1 and β_2 and σ for different values of α

9.2 Belgium telephone call data

The real data set for our second example is from the Belgian Statistical Survey by the Ministry of Economy and contains the total number (in tens of millions) of international phone calls made in a year from 1950 to 1973. Our data set actually has a scalar multiple of 10 compared to the data set used by the original paper, So each of our estimates and the graphs may be a scale of 10. For this data, the independent variable x is the year, and the dependent variable y is the number of telephone calls. Obviously, we see that the points are not too smooth and the points corresponding from 1964 to 1969 show a completely different

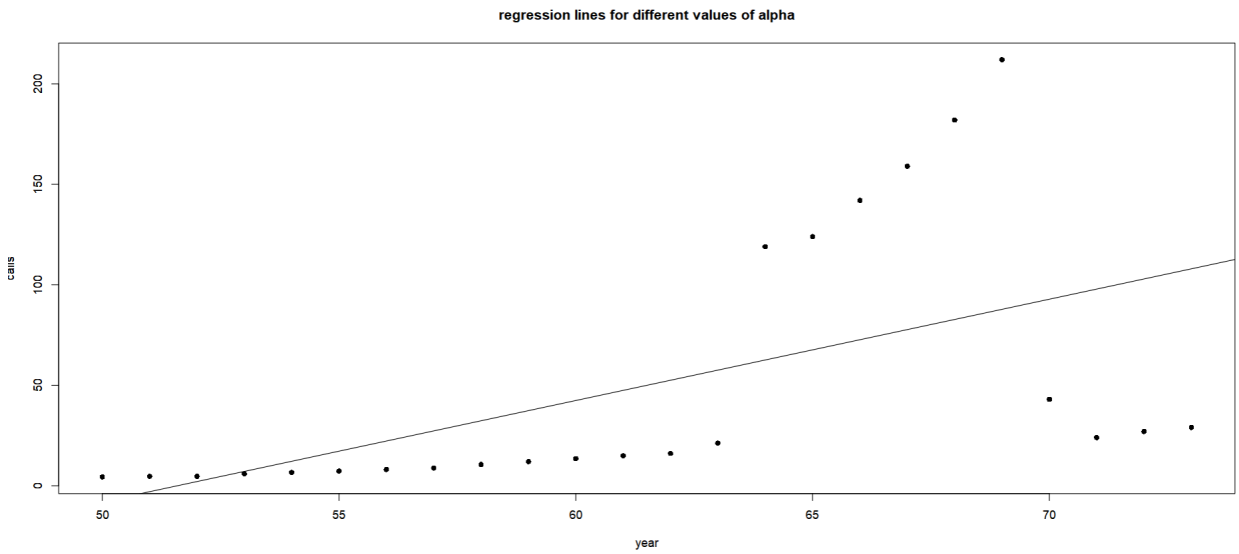

Figure 8: scatter plot with least square regression line

pattern, hence behaving like outliers. As we have seen from the plot that these six-year (1964 to 1969) have a very high amount of telephone calls compared to the other years. So they are significantly different from the other data points, and these points are considered outliers.

As the picture shows that the least square regression line is heavily affected by the outliers and fits the data too worse making the estimation worthless. After that, we will follow the theory of the MDPD estimator to verify our analytical justification from a full practical point of view.

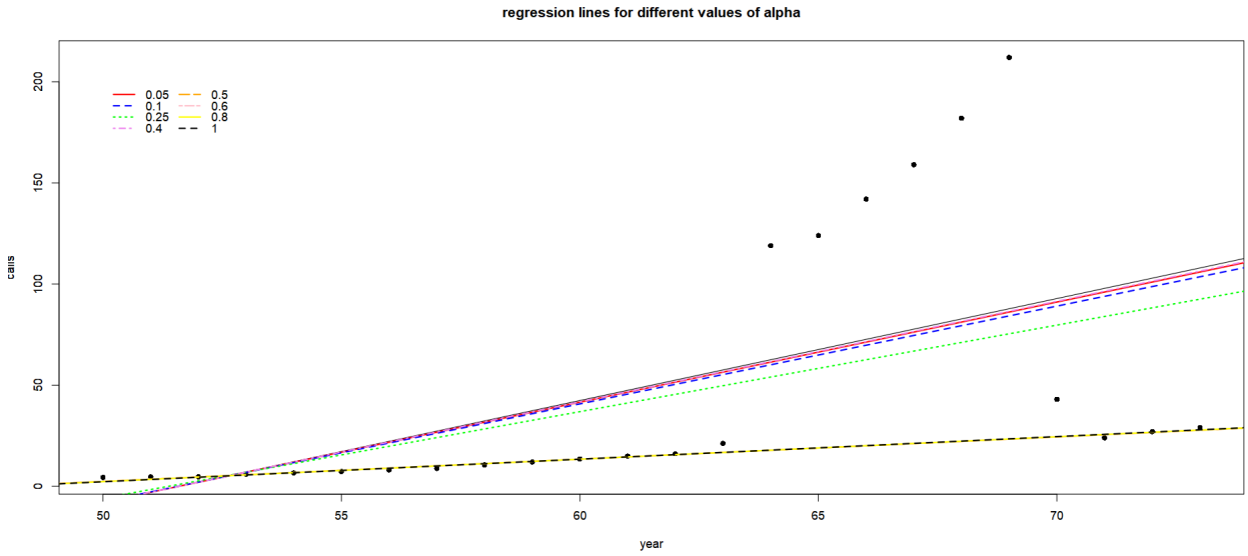

Figure 9: scatter plot with least square and DPD regression line

In this case, the breakdown point is

$$\frac{[n-p]/2+1}{n} = \frac{[24-2]/2+1}{24} = 0.48$$

So after that, all the estimates are highly robust and before that, all the data points are highly efficient. So we can choose any value of $\alpha > 0.50$. So any estimate greater than the above cutoff value is highly preferable from a robust point of view. Finally, we will show the estimate of the parameters in its numerical figure.

	0	0.05	0.1	0.25	0.4	0.5	0.6	0.8	1
β_1	-260.05	-255.30	-249.35	-219.77	-256.38	-52.56	-52.75	-53.14	-53.57
β_2	5.041	4.94	4.83	4.27	4.96	1.09	1.10	1.10	1.11
σ	56.22	54.01	54.06	52.91	53.98	1.12	1.14	1.19	1.23

Table 3: Different values of β_1 and β_2 and σ for different values of α

9.3 Salinity data

Finally, as an example of the multiple regression model with masking effects, we consider the "Salinity data" that were originally presented by Ruppert and Carroll (1980). The data set contains measurements of the salt concentration of the water and the river discharge taken in North Carolina's Pamlico Sound. Rousseeuw and Leroy (1987) consider this data as multiple linear models with salinity as the dependent variable and the independents variables being salinity lagged by two weeks (x1), the number of biweekly periods elapsed since the beginning of the spring season (x2), and the volume of river discharge into the sound (x3). But here in this data also we will also see some sort of outliers (not graphically since the dimension is more than three) in the residual analysis. It is not obvious which points are residuals and which are not is not so clear from the above figure. All the points look haphazard in layman's language. So the model may be very well fitted. But to see how much it fits the data we have to study the residuals further and we will not do that. We will fit different models corresponding to different α values.

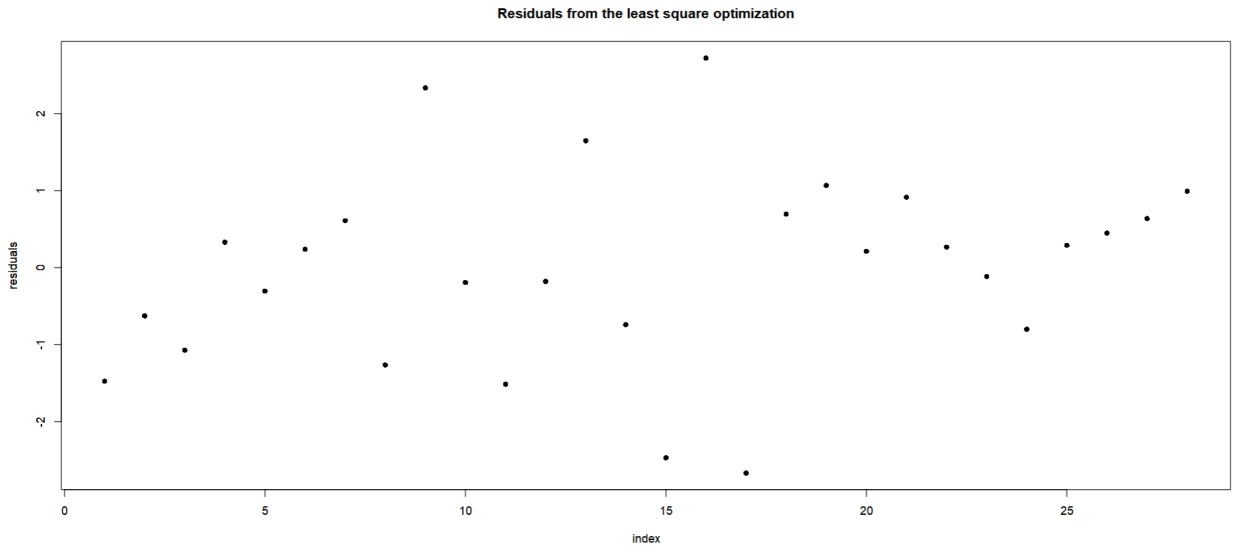

Figure 10: residual plot with least square regression

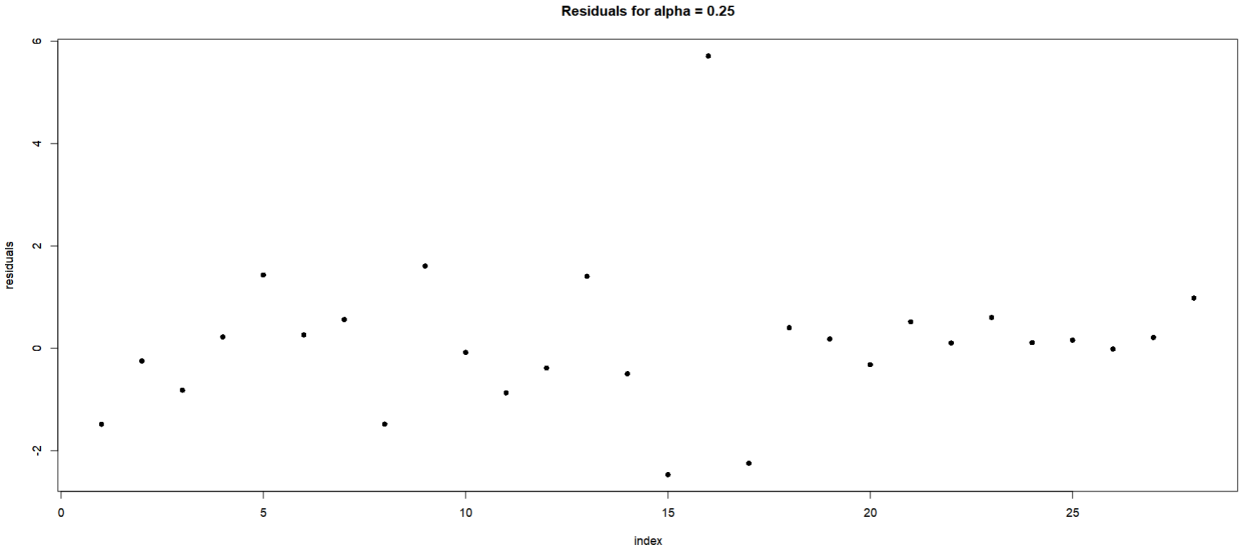

Figure 11: residual plot with DPD estimator with α =0.25

In this case, the breakdown point is

$$\frac{[n-p]/2}{n} = \frac{[28-4]/2+1}{28} = 0.46$$

So after that, all the estimates are highly robust and before that, all the data points are highly efficient. So we can choose any value of $\alpha > 0.46$. So any estimate greater than the above cutoff value is highly preferable from a robust point of view.

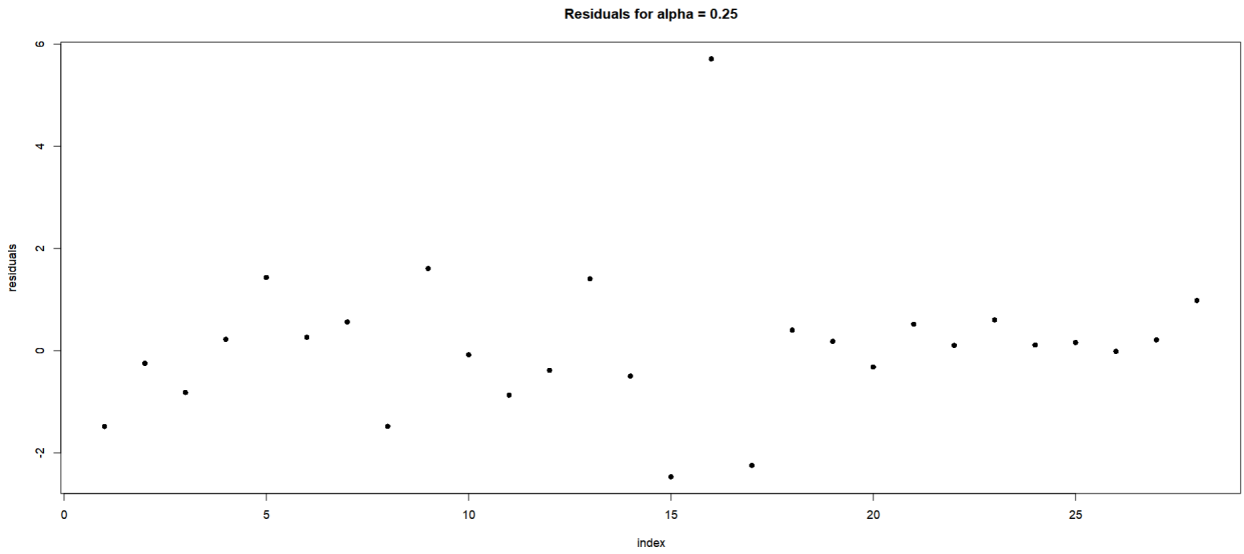

Figure 12: residual plot with DPD estimator with α =0.50

Now we see that both the plots are indicating that point index 16 is actually an outlier point and if we choose the value of $\alpha = 0.5$ then things become more realistic or the outlier is much more distinguishable from the others. Hence finally choose the value of $\alpha = 0.50$

	0	0.05	0.1	0.25	0.4	0.5	0.6	0.8	1
β_1	9.59	9.93	10.48	18.01	9.873	18.38	18.38	11.24	1.15
β_2	0.77	0.77	0.77	0.71	0.77	0.72	0.72	0.78	1.12
β_3	-0.02	-0.03	-0.04	-0.17	-0.03	-0.19	-0.19	-0.04	0.22
β_4	-0.29	-0.30	-0.32	-0.61	-0.30	-0.62	-0.62	-0.36	-0.09
σ	1.33	1.22	1.21	0.96	1.22	0.86	0.83	0.80	1.42

Table 4: Different values of β_1 ,β_2 ,$\beta_3$and β_4 and σ for different values of α

10 Conclusion

The most obvious parameter estimation for the normal linear regression is the least square fitting but it does not have the desirable property of robustness. It has the highest efficiency but it is too much dependent on the entire data. Slightly changing the data causes a large change in the least square estimate even if not in the case of multicollinearity.

But, here we suggest that sacrificing a small amount of efficiency gives us an estimate which is much more robust than the least square estimate and how much we will sacrifice the efficiency is in our hands, so overall we proposed a better estimator from a robust point of view.

So we first developed the theory of a proposed estimator named Minimum Density Power Divergence Estimator. We first derived its asymptotic behavior that the rate of convergence

for this estimator and find that this estimator has a rate of convergence the same as that of the MLE estimator.

So the next important criterion for any estimator is the limiting variance or the Asymptotic Relative Efficiency(ARE). We compared our estimator to the corresponding MLE and we find out that the ARE decreases as increasing α. So it causes the estimator not only be inefficient but also useless to predict and situation.

Next, we looked into the robustness behavior. For this, we study the Influence function. Here if it goes to infinity we say that the estimator is non-robust. If it is bounded then the estimator will be highly robust. We see that our function is bounded and the MLE is unbounded. So our estimator has a highly robust behavior.

Finally, we looked into the case of simulated data and also the case of real data and find that the theoretical results are perfectly applicable to our case. For some cases, we choose a low α value and for some cases, we have to choose a high α value by looking at the fitting quality of the estimator.

11 Contribution

- Prithwijit Ghosh 60% Coding + paper finding + 20% report
- Gagan Deep 40% report + 40% Presentation + 20%coding
- Sarvesh Singh Khushwaha 20%report +20%presentation+10%coding
- Gaurav 10% coding + 20% report +40% presentation

12 Acknowledgement

We take this opportunity to heartily thank our supervisor Prof. Arnab Hazra for his valuable feedback and constant guidance on this project.