# Strategy-Proof Estimators for Simple Regression[*]

*Authors*: Javier Perote and Juan Perote-Peña.

*Author 1*: Javier Perote

*Affiliation:* Facultad de C.C. Jurídicas y Sociales, Universidad Rey Juan Carlos.

*Address*: Universidad Rey Juan Carlos, Facultad de C.C. Jurídicas y Sociales, Campus de Vicálvaro, P$^o$de los Astilleros s/n, 28032 Madrid, Spain.

*Telephone*: +34-913019923.

*E-mail*: perote@fcjs.urjc.es

*Author 2*: Juan Perote-Peña

A*ffiliation:* Departamento de Economia y Empresa, Universidad Pablo de Olavide de Sevilla y Fundación Centro de Estudios Andaluces (CentrA, internal researcher in the Microeconomics group).

*Running title*: Solidarity versus Flexibility.

*Address for manuscript correspondence*:

Juan Perote Peña, Departamento de Economía y Empresa, Universidad Pablo de Olavide de Sevilla, Carretera de Utrera, Km. 1, 41013, Sevilla, Spain.

*Telephone*: +34-954349187.

*E-mail*: jperpen@dee.upo.es

*Fax*: +34-954349339

[*] We thank Ashley Piggins and an anonymous referee for very helpful comments on the first draft and Fundación Centro de Estudios Andaluces (CentrA) for financial support.

*Abstract*

In this paper we propose a whole class of estimators (*clockwise repeated median estimators* or CRM) for the simple regression model that are immune to manipulation by the agents generating the data. Although strategic considerations affecting the stability of the estimated parameters in regression models have already been studied (the Lucas critique), few efforts have been made to design estimators that are incentive compatible. We find that some well-known robust estimators proposed in the literature like the resistant line method are included in our family. Finally, we also undertake a Monte Carlo study to compare the distribution of some estimators that are robust to data manipulation with the OLS estimators under different scenarios.

*Keywords*: strategy-proofness, single-peaked preferences, robust regression, data contamination.

*JEL classification numbers*: D78, C13.

# 1 Introduction

The search for robustness of linear regression estimators has been traditionally motivated by the existence of contaminated samples or the potential presence of outliers. In this paper we interpret sample contamination in terms of sample manipulation. Therefore we consider the problem of estimating a linear regression in which the values of the dependent variable are provided by strategic agents endowed with private non-observable information whilst the values for the regressors are verifiable public information. We shall assume that every agent that is behind each "observation" in the sample is better off the closer the prediction obtained with the regression is from the dependent variable's true value.

Therefore, the agents underlying the data might have an incentive to report a false (unverifiable) information about the dependent variable in order to obtain a better prediction for themselves at the cost of biasing the regression and imposing other costs on the remaining agents. Hence, the agents behind the observations have an incentive to analyze the regression method used by the econometrician and try to find scope for obtaining a better result by exploiting their private information when their objective is not being a true outlier.

Notice that whenever it is possible to use a non-linear regression method, that allows for a potentially different estimator for each agent, the truthful revelation issue is no longer a problem. The prediction for each observation can actually be the declared value of the dependent variable and every agent behind each observation would be as better as possible, so that there will be no need to lie whatsoever. Moreover, we could also use any jack-knife estimator in the literature that considers the subsample obtained by excluding just one observation to generate the prediction for each agent (the excluded one). Since the information reported by any agent will therefore not be used to produce her own prediction (only to generate the others' predictions), the estimation method does not give incentives to lie either. Our incentive compatibility problem begins when predictions must follow from a single linear regression line (a single estimator for all the agents that has a somehow "public good" nature).

There exist different contexts where the reliability of the data is objectionable because the data could come from surveys composed by agents interested in not being perceived as real outliers if the estimation results could be used in the future to change the economic situation of the agents that generate the sample, for example. Think, for example, of the case of estimating the productivity of a set of divisions of a bureaucracy that in principle share the same average technological characteristics but there is a random shock that could either increase a division's true productivity (good luck) or drop it below the average (bad luck). If the dependent variable observation of each division is private information, a division that had, say, bad luck could fear being treated as a low productivity outlier and thus could be penalized in the future. For example, the division may have a reduced future budget, or could face a higher risk of closure. Therefore, if OLS regression is used, the division could be tempted to declare an even smaller output in order to bias the regression in such a way that the prediction for herself is closer to her true response variable value. The false information reported by the division therefore contaminates the sample. The division could be interested in bringing the regression line closer to her true productivity because a future possible inspection would find out that the division's true value is not too big when compared to the predicted productivity using the regression line as a reference[1]. Therefore, exaggeration of both bad luck and good luck can be profitable in many contexts and the data reported by the agents will not be the true sample. Nevertheless, the compatibility of individual incentives and the social planner estimation can be achieved by designing estimators which are linear regression robust estimates immune to strategic manipulation.

In *Section 2*, the theoretical model is introduced and the definitions about the kind of strategic contamination we are interested in avoiding are established. We find a family of estimators for the simple regression model called *clockwise repeated median estimators* (CRM) that happen to be resistant to some kind of strategic data contamination. In particular, the family encompasses some robust estimation methods like the resistant line one (Tukey, 1970/71) or variations on other well-known methods. We also prove that OLS and other estimators are not immune to strategic contamination.

*Section 3* deals with some Monte Carlo experiments that simulate the behavior of the OLS and the CRM estimates under two different scenarios. First, we compare the distributions obtained with OLS regression applied to contaminated data with the estimates obtained with some of our CRM estimators applied to the clean true sample. Then, we proceed to check the efficiency loss that occur when using some of our estimators with respect to OLS when applied to the true sample. We conclude (*Section 4*) that if the problem of strategic contamination is severe enough (for example, when the true observations in the sample cannot be easily verified), CRM estimators can provide better estimates than other methods that are not robust to data manipulation like OLS estimators. A relatively small loss of consistency can be compensated with the accuracy of non-contaminated data.

[1] Notice that the true value of the response variable for each division is fixed and cannot change, although the planner or the econometrician cannot directly observe it.

# 2 The model

Let us assume the following simple regression data generating process:

$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ for } i = 1, ..., n. \tag{1}$$

where $n \geq 3$ is the number of observations or sample size. We denote the set of names of the observations as $N = \{1, ..., n\}$. Variable $x_i$ is the explanatory variable (real), $y_i$ is the response or dependent variable and $e_i$ is an error term or random shock that is i.i.d. and normally distributed with zero mean and variance $\sigma^2$, i.e.: $e_i \sim N(0, \sigma)$.[2] Traditionally econometricians make the implicit assumption that the sample generated by (1) is fully observable and therefore the problem consists in estimating the vector of unknown parameters $\beta' = (\beta_0, \beta_1)$ from the data matrices:

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ and } Y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \tag{2}$$

By using an appropriate regression estimator $\widehat{\beta}' = (\widehat{\beta}_0, \widehat{\beta}_1) = T(X, Y)$ that is a function $T$ of the sample $(X, Y)$, we obtain the regression coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Given these estimated regression coefficients, we can obtain the predicted or estimated values of the response variable $\widehat{y_i}$ for each observation $x_i$ using the function:

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i. \tag{3}$$

The residual $\widehat{e_i}$ is defined as the difference between the true response variable for the $i$th observation and the predicted one, i.e., $\widehat{e_i} = y_i - \widehat{y_i}$. Regression analysis deals with the problem of designing regression estimators that are well-behaved mainly in the sense of providing good fits, i.e., estimators $\widehat{\beta}' = T(X, Y)$ such that for every possible admissible sample $(X, Y)$ generate small residuals or "reasonably good fits". The most commonly used regression estimator is, of course, the ordinary least squares estimator (OLS) that minimize the sum of the squared residuals for every sample. Nevertheless, in some cases we are interested in estimators that provide not only reasonably good fits but also satisfy other desirable criteria.

[2] Our results do not actually depend on the distribution of the disturbance. We assume the most typical econometric model to keep things simple.

For example, if there is a risk of the sample being "contaminated", i.e., it is possible that some of the observations do not really come from the data generating process in (1), usual estimators like OLS are very sensitive to the value of the observations and the introduction of this kind of "outliers" can heavily affect the regression results. When the reason for the presence of outliers is random typing or information managing errors, robust estimators that are not so sensitive to data contamination are proposed to cope with the problem (see Rousseeuw et al. (1987) for an excellent survey on the topic). In this paper we deal with a different kind of data contamination created by strategic behavior: imagine that the observations $y_i$ are a measure of individual attitudes towards some key issue that are revealed from some agents potentially involved in the results, whereas values of $x_i$ come from a verifiable source immune to contamination. Let us illustrate our problem by means of some examples.

Consider first the problem of a monopolist national trade union in any industrial sector like steel, for example, that has to decide which hourly wage to set in the market. The trade union has to collect information about the actual pre-wage bill expected profitability of each of the individual firms' unions in the sector $(y_i)$ together with the number of hours worked in each firm $(x_i)$, aggregate the information and produce a single sectorial wage per hour $(\widehat{\beta}_1)$ and a fixed individual benefit (independent of the number of hours worked in each firm, $\widehat{\beta}_0$). The problem is that although the number of hours worked in each firm is public reliable information, the pre-wage bill expected profit each firm has is a private information owned by the managers and workers inside each individual firm, since the information will only become available after the decision about the common wage is taken.

Imagine that the central union wants to minimize the sum of the squared residuals and therefore uses OLS to decide the common wage $\widehat{\beta}_1$ and fixed benefit $\widehat{\beta}_0$. The workers in those firms that are less profitable, efficient or those that had bad luck will typically show a negative residual $\widehat{e}_i = y_i - \widehat{y}_i = y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) < 0$, and the final wage bill for the firm will exceed the individual pre-wage bill profit of the firm, that will have to fire some workers or increase the risk of closing it down. In order to lower the wage bill in accordance with the financial possibilities of the firm, the workers can report a pre-wage bill profit even smaller and lower the common wage down to the actual average productivity of labor within the firm by pushing the regression line closer to their true observation $(x_i, y_i)$. Identically, the workers in those firms that are more profitable, efficient or lucky will have positive and big residuals, so that there is still scope to increase the hourly common wage and fixed benefit closer to real average productivity of labor in the firm without fearing being sacked or increasing the risk of closure, so the workers in these firms will have an incentive to exaggerate their own profitability in order to achieve higher wages. The central union that uses OLS when deciding the common wage to set in the sector cannot expect the workers' representatives within the individual firms to behave truthfully.

A possibility to induce truthful revelation would be to use the jack-knife estimator to calculate the residual for each firm $i \in N$ (OLS with the subsample in which the $i$th observation is deleted), but doing this will allow for $n$ potentially different wages and benefits in the market, which will cause a costly distortion in the market that cannot be sustained within a market economy where the unavoidable re-allocation of inputs will trigger a single wage. The need to fix a unique common wage for all firms has therefore a "public good" nature that prevents the union (or the planner) to use a non-linear regression. We assume therefore that there are important transaction costs that constrain the designer to use a single linear regression to generate the residuals.

Other example can be the design of a simple linear tax schedule on income levels in order to implement a simple pay-as-you-earn tax collecting system. The final tax the tax-payers will pay is independent of the fraction of their monthly income that will be diverted to the Treasure, but the tax-payers' preferences for liquidity and risk aversion define their preferences on the monthly average tax and the most-preferred average tax $(y_i)$ is likely to be positively related with individual income $(x_i)$ under a progressive tax scheme. The need for a simple and comprehensive linear tax schedule excludes jack-knife-type estimators.

A different example could be a questionnaire among workers in the same factory that earn extra income from farming, for example, about the total individual income they earn $(y_i)$. Let $x_i$ be a measure of some relevant characteristic like their ages. We could assume that equation (1) is generating the true observations, but since the response variable $y_i$ is private information of each individual, the statistician has to rely on both the honesty of the workers and good questionnaire design. Nevertheless, imagine that the workers fear that the data could be analyzed to detect outliers and sold to the Tax Inspection agency and they could be penalized in the future or they could fear to be discriminated as lazy workers. Individuals have an incentive to not being considered as true outliers if they had a good year or a bad one and to hide themselves among the majority if they think that the data will be aggregated by using estimators too sensitive to individual observations like OLS.

Finally, consider the problem of estimating the average productivity of a set of divisions within a bureaucracy or a big corporation when there is no reliable measure of the output other than asking the divisions to report their own levels of activity. Variable $y_i$ might be the unverifiable measure of the output of division $i$ and $x_i$ the amount of the variable input used (say, the budget available to division $i$ or the number of workers). Again, there is an incentive not to be suspect of being a true outlier. Outliers are typically observations with too large residuals, so division $i$ would have an incentive to lie and report an output level such that the predicted one is as close as possible to his own true $y_i$ (in other words, to minimize the residual $\widehat{e_i}$) if OLS is supposed to be used as the regression estimator to generate the predictions. Since every division has an incentive to minimize his own residual regardless of the higher residuals imposed on others, the revealed $y_i's$ would be typically contaminated by the strategic behavior of the agents behind the data and the regression coefficients will lose their "good fit" properties (a bias could be introduced and, depending on the importance of the problem, the regression could be meaningless). In *Section 3* we perform a Monte Carlo study of OLS estimators when the sample is contaminated by strategic agents.

What can we do in the case of strategic contamination of the data? If the problem is severe enough, we still can design the estimators to be immune to likely forms of manipulation, at a cost in terms of other properties that will be lost (less consistency or asymptotic bias, difficulty of calculation, etc.). We propose a property of strategic non-manipulability of the data well-known in the social choice and incentives literature that amounts to imposing that reporting the true value of the response variable for each $i$ is a dominant strategy for the agent behind observation $i$ given the estimator used. No agent behind any observation will have any incentive to lie in order to reduce the obtained residual for every true sample. This requirement amounts to *strategy-proofness* when translated to the regression context. Strategy-proofness is a well-known incentive compatibility property (see, for example, Gibbard (1973)). Let us be more explicit about this property. We denote as agent $i \in \{1, ..., n\}$ the rational agent behind observation $i$. If $E$ denotes the real line, given any true value $y_i \in E$, each agent $i$ will have a complete and continuous preference relation $R_i^{y_i}$ on $E$ (the predictions space). Let $P_i^{y_i}$ be the asymmetric part of $R_i^{y_i}$, $\forall i \in \{1, ..., n\}$ . In what follows, we must distinguish between the unobservable true response variables for each $i$ and the revealed, declared or reported ones. Let us denote as $\widetilde{y_i}$ the response variable actually reported to the researcher by the agent behind the $i$th observation. Estimators must be defined on the agents' reported response variable, not on the true values, so potential strategic contamination of the sample obliges us to define an estimator as a function of the reported rather than the true values: $\widehat{\beta}' = (\widehat{\beta}_0, \widehat{\beta}_1) = T(X, \widetilde{Y})$. We also assume a particular case of manipulation: agents are interested in not being suspect of being potential true outliers. We represent this fact by defining a particular form to the preferences of the agent behind each observation on the predicted value with respect to the true one.[3]

**Definition 1** *Agent $i$ with true response value $y_i$ has* single-peaked preferences $R_i^{y_i}$ *when the following holds: (i). $y_i P_i^{y_i} v$ $\forall v \in E$, $v \neq y_i$. (ii). $\forall v, \overline{v} > 0$, $\overline{v} > v \rightarrow (y_i + v) P_i^{y_i} (y_i + \overline{v})$ and $(y_i - v) P_i^{y_i} (y_i - \overline{v})$.*

Single-peaked preferences were first introduced by Black (1958) and their strategic properties have been extensively analyzed (see Barberà et al. (1994), for example). Let us denote as $\Re^{y_i}$ the domain of every single-peaked preference for agent $i \in \{1, ..., n\}$ with true value $y_i \in E$. Notice that for the agents underlying the observations, the best possible situation is that of being predicted their true values (that is, (i) in the definition above) and the further away the prediction is with respect to the true value, the worse off the agent is (part (ii) in *Definition 1*). There is still scope for different sensibilities when judging a positive residual with a negative one. From now on, and abusing notation, we shall consider $Z = (X, \widetilde{Y})$ to be such that $X$ could be restricted to some subset of all possible values and therefore, the definitions below refer to the restricted domain for $Z$ considered in the problem. Besides, we shall make use of the notation $(y_i, Y_{-i}) = Y$.

**Definition 2** *Regression estimator* $\widehat{\beta}' = (\widehat{\beta}_0, \widehat{\beta}_1) = T(X, \widetilde{Y}) = T((X, \widetilde{y}_i, \widetilde{Y}_{-i}))$ *is* manipulable at sample $(X, \widetilde{Y}) \in Z$ *by observation* $i \in \{1, ..., n\}$ *if* $\exists R_i^{\widetilde{y}_i} \in \Re^{\widetilde{y}_i}$, $\exists \overline{y}_i \in E$, $(\overline{y}_i \neq \widetilde{y}_i)$ *such that*

$$\left[\widehat{\beta}_0(X, \overline{y}_i, \widetilde{Y}_{-i}) + \widehat{\beta}_1(X, \overline{y}_i, \widetilde{Y}_{-i}) x_i\right] P_i^{\widetilde{y}_i} \left[\widehat{\beta}_0(X, \widetilde{y}_i, \widetilde{Y}_{-i}) + \widehat{\beta}_1(X, \widetilde{y}_i, \widetilde{Y}_{-i}) x_i\right].$$

**Definition 3** *Regression estimator* $\widehat{\beta}' = (\widehat{\beta}_0, \widehat{\beta}_1) = T(X, \widetilde{Y}) = T((X, \widetilde{y}_i, \widetilde{Y}_{-i}))$ *is* strategy-proof *if it is not manipulable at any sample* $(X, \widetilde{Y}) \in Z$ *for any observation* $i \in \{1, ..., n\}$.

Strategy-proof estimators are also called *non-manipulable*. A strategy-proof regression estimator leaves no gain in declaring a false response variable value $(\widetilde{y}_i \neq y_i)$ for no agent behind any observation $i \in \{1, .., n\}$. A rational

[3] It might well be the case of the true value being observed later or at a closer random investigation and there is no cost in reporting over-estimated or under-estimated values (there is always the possibility of claiming to have reported involuntary errors).

agent can only (weakly) worsen his own prediction for any single-peaked preferences on the prediction space by reporting false information.

In this paper we propose a whole family of strategy-proof estimators robust to individual strategic manipulation of the response variable called clockwise repeated median estimators (CRM). Moreover, every CRM estimator is such that the regression line always pass through two different observations for any admissible sample. Nevertheless, CRM estimators do not exhaust the class of strategy-proof estimators. Later, we shall prove that common estimators in the literature leave scope for strategic manipulation. We shall see that OLS and some well-known estimators that are robust to non-strategic contamination such as the *least median of squares estimator* (Rousseeuw (1984)), Theil's (1950) estimator and Siegel's (1982) repeated median estimators are not strategy-proof. The first example of non-manipulable estimator we propose works for the case of $\beta_0 = 0$ and $x_i > 0 \; \forall i \in \{1, ..., n\}$ and amounts to an extension of the Median Voter Theorem. Let us define the Median Voter (MV) estimator as:

$$\begin{aligned} \widehat{\beta}_1 &= med\left\{\frac{\widetilde{y}_i}{x_i}\right\} \\ \widehat{\beta}_0 &= med_{i\in N}\left(\widetilde{y}_i - \widehat{\beta}_1 x_i\right) \end{aligned} \tag{4}$$

We can easily establish the following result:

**Proposition 1** *The MV estimator is strategy-proof for all admissible samples $Z'$ with $x_i > 0 \; \forall i \in N$.*

**Proof.** Let us take any sample $(X, \widetilde{Y})$. We define the variable $w(i, \widetilde{y}_i, x_i) = \dfrac{\widetilde{y}_i}{x_i} \; \forall i \in N, \; \forall \widetilde{y}_i \in E, \; \forall x_i \in E_{++}$. For any single-peaked preferences $R_i^{y_i}$ for all $i \in N$, we can define a different single-peaked preferences $\overline{R}_i^{y_i}$ such that for all $v, v' \in E, \;\; w(i, \widehat{y}_i, x_i)\overline{R}_i^{y_i} w(i, \widehat{y}'_i, x_i) \longleftrightarrow (\widehat{y}_i) R_i^{y_i} (\widehat{y}'_i)$. Since $\widehat{\beta}_0 = 0$ for every sample, we can interpret each reported individual slope $w(i, \widetilde{y}_i, x_i)$ as the reported "peak" of each agent's single-peaked preferences $\overline{R}_i^{y_i}$. Notice that by construction, $\left(\dfrac{y_i}{x_i}\right) \overline{P}_i^{y_i} \left(\dfrac{\widetilde{y}_i}{x_i}\right) \;\; \forall \widetilde{y}_i \in E, \; \forall x_i \in E_{++}$. Then, $\widehat{\beta}$ MV is strategy-proof if and only if $\left[med\left\{\dfrac{y_i}{x_i}\right\} x_i\right] R_i^{y_i} \left[med\left\{\dfrac{\widetilde{y}_i}{x_i}\right\} x_i\right]$ $\forall i \in N, \; \forall \widetilde{y}_i, y_i \in E, \; \forall x_i \in E_{++}$. But by construction this implies that $\dfrac{\left[med\left\{\dfrac{y_i}{x_i}\right\} x_i\right]}{x_i} \overline{R}_i^{y_i} \dfrac{\left[med\left\{\dfrac{\widetilde{y}_i}{x_i}\right\} x_i\right]}{x_i} \; \forall i \in N, \; \forall \widetilde{y}_i, y_i \in E, \; \forall x_i \in E_{++}$, or simplifying terms: $med\left\{\dfrac{y_i}{x_i}\right\} \overline{R}_i^{y_i} med\left\{\dfrac{\widetilde{y}_i}{x_i}\right\} \; \forall i \in N, \; \forall \dfrac{y_i}{x_i}, \dfrac{\widetilde{y}_i}{x_i} \in E$, given any admissible $X$. Then, we know that choosing the median on a single dimension is a strategy-proof allocation method (see Moulin (1980)), so $\widehat{\beta}$ MV must be strategy-proof. ■

The MV estimator has nevertheless limited interest, since the admissible samples for which it works exclude negative values for the $x_i$'s and it always imposes a zero estimate for the intercept. We shall introduce now a class of estimators that are strategy-proof for almost all samples and for every simple regression. The relevant samples for calculating any CRM estimator are those $\overline{Z} = (X, \widetilde{Y})$ such that $x_i \neq x_j$ for all $i, j \in N$. Let us now define this class of *clockwise repeated median* (CRM) estimators: First, each member of the class is parameterized by two fixed sets of names of observations: $S, S' \subseteq N = \{1, ..., n\}$ such that either $S \cap S' = \emptyset$ or $S \subseteq S'$. Then, we need to calculate the *clockwise angle* of any pair of declared observations $i, j \in \{1, ...n\}$ in the sample, $CWA((x_i, \widetilde{y}_i), (x_j, \widetilde{y}_j))$, defined as: $\forall (x_i, \widetilde{y}_i), (x_j, \widetilde{y}_j) \in (X, \widetilde{Y})$,

$$CWA((x_i, \widetilde{y}_i), (x_j, \widetilde{y}_j)) = \pi + sign(x_j - x_i)\frac{\pi}{2} + sign\left(\frac{\widetilde{y}_j - \widetilde{y}_i}{x_j - x_i}\right)\left|\arctan\left(\frac{\widetilde{y}_j - \widetilde{y}_i}{x_j - x_i}\right)\right|. \tag{5}$$

Then, we can define the *directing angle*, $DA(X, \widetilde{Y})$, defined such as $\forall (X, \widetilde{Y}) \in \overline{Z}$:

$$DA(X, \widetilde{Y}) = med_{i \in S}\ med_{\substack{j \in S' \\ j \neq i}}\ CWA((x_i, \widetilde{y}_i), (x_j, \widetilde{y}_j)). \tag{6}$$

And finally, the regression estimator is obtained with the following formula:[4]

$$\begin{aligned} \widehat{\beta}_1 &= \tan\left[DA(X, \widetilde{Y}) - \pi - \frac{\pi}{2} sign(DA(X, \widetilde{Y}) - \pi)\right] \\ \widehat{\beta}_0 &= med_{i \in S}\ \left(\widetilde{y}_i - \widehat{\beta}_1 x_i\right) \end{aligned} \tag{7}$$

Although the description of the estimator seems complicated, it has a clear intuition. First, the estimate for the slope is generated as follows: for every agent or observation in $S$, we take each observation $i \in S$, say $(x_i, \widetilde{y}_i) \in (X, \widetilde{Y})$, find the straight line that passes through $(x_i, \widetilde{y}_i)$ and any other observation in the set $S' \subseteq \{1, ..., n\}$ and rank any other observation from the one with the smallest clockwise angle to the one with the highest starting from 12 o'clock until we exhaust all other observations. Then, we

[4] Notice that $DA(X, \widetilde{Y}) - \pi \neq 0$, by construction for all admissible samples $(X, \widetilde{Y}) \in Z$.

find the median angle[5] of the ranked angles for $(x_i, \widetilde{y}_i)$, and again the median of the medians of the angles for all observations in the set $S$, which we call the directing angle. Finally, we transform the directing angle into a slope by means of the tangent function and this will be estimator $\widehat{\beta}_1$. The estimator for $\beta_0$ is the median of the intercepts with the $y$-axis of the projected lines that pass through each observation in set $S \subseteq N$ and share the same slope $\widehat{\beta}_1$. The class of CRM estimators are therefore parameterized by the sets $(S, S')$.

Let us analyze the example shown in *Figure 1.* We have a sample $(X, \widetilde{Y})$ composed by four observations: $(x_1, \widetilde{y}_1), (x_2, \widetilde{y}_2), (x_3, \widetilde{y}_3), (x_4, \widetilde{y}_4)$. Let us consider the CRM estimator defined by $S = S' = \{1, 2, 3, 4\}$. First, we start by observation $(x_1, \widetilde{y}_1) \in S$ and find $CWA((x_1, \widetilde{y}_1), (x_2, \widetilde{y}_2))$. The expression (5) above represents the angle shown as the dashed sector in *Figure 1.1.* Analogously, *Figure 1.2.* shows the angle $CWA((x_1, \widetilde{y}_1), (x_4, \widetilde{y}_4))$. To find the directing angle, we first need to find $med_{j\in S'} CWA((x_1, \widetilde{y}_1), (x_j, \widetilde{y}_j)) = CWA((x_1, \widetilde{y}_1), (x_2, \widetilde{y}_2))$. Now, we proceed likewise taking as the reference observation $(x_2, \widetilde{y}_2)$ and easily find that $med_{j\in S'} CWA((x_2, \widetilde{y}_2), (x_j, \widetilde{y}_j)) = CWA((x_2, \widetilde{y}_2), (x_1, \widetilde{y}_1))$. Now, we take the third observation and get

$med_{j\in S'} CWA((x_3, \widetilde{y}_3), (x_j, \widetilde{y}_j)) = CWA((x_3, \widetilde{y}_3), (x_4, \widetilde{y}_4))$. Finally, taking $(x_4, \widetilde{y}_4)$, we calculate $med_{j\in S'} CWA((x_4, \widetilde{y}_4), (x_j, \widetilde{y}_j)) = CWA((x_4, \widetilde{y}_4), (x_3, \widetilde{y}_3))$. The four angles (one for each observation) which are candidates to be the directing angle are depicted as pointing arrows in *Figure 1.3.* Now, we can apply expression (6) to find the directing angle using the repeated median, so $DA(X, \widetilde{Y}) = med_{i\in S}\; med_{\substack{j\in S' \\ j\neq i}}\; CWA((x_i, \widetilde{y}_i), (x_j, \widetilde{y}_j)) =$

$$= med \left\{ \begin{array}{c} CWA((x_1, \widetilde{y}_1), (x_2, \widetilde{y}_2)), CWA((x_2, \widetilde{y}_2), (x_1, \widetilde{y}_1)), \\ CWA((x_3, \widetilde{y}_3), (x_4, \widetilde{y}_4)), CWA((x_4, \widetilde{y}_4), (x_3, \widetilde{y}_3)) \end{array} \right\} =$$

$= CWA((x_1, \widetilde{y}_1), (x_2, \widetilde{y}_2))$. Notice that we have decided to define the median of an even number of angles as the maximum median. Now, we must transform the directing angle into a slope by using the formula to find $\widehat{\beta}_1$ in (7), that simply undo the angle formula to find the corresponding slope, which is depicted in *Figure 1.4.* The dashed straight line embodies the slope of the regression using this CRM estimator. The estimate for the intercept is calculated from (7) by projecting the slope passing through each observation to find the median of the intercepts. Clearly, the dashed line passing through both $(x_1, \widetilde{y}_1)$ and $(x_2, \widetilde{y}_2)$ is the regression line. Observe that it must always be the case of the regression line passing through two observations in the sample. The clockwise angle supporting the slope of the regression line is the

[5] When calculating the median of an even number of observations it can be used either the bigger or the smaller of the two median observations, but not the average if we want to preserve strategy-proofness.

directing angle.

[Insert *Figure 1* about here]

Let us analyze some members of the class of CRM estimators:

If $S = S' = N$, we obtain

$$DA(X, \widetilde{Y}) = med_{i \in N}\ med_{j \in N \backslash \{i\}}\ CWA((x_i, \widetilde{y}_i), (x_j, \widetilde{y}_j)), \tag{8}$$

The resulting estimator is an extension of the *repeated median estimator* defined in Siegel (1982) when the medians are taken on the angles defined by the slopes, not on the slopes themselves. The breakdown point of this estimator is the biggest possible, 50%, i.e., the estimator remains bounded unless at least half of the observations' response variables go to infinity.

Given any $h \in N$, if $S = N \backslash \{h\}$ and $S' = \{h\}$, we obtain

$$DA(X, \widetilde{Y}) = med_{i \in N \backslash \{h\}}\ CWA((x_i, \widetilde{y}_i), (x_h, \widetilde{y}_h)), \tag{9}$$

which is a clockwise median extension of the *median star estimator* (Simon (1986)) when defined taking as the reference observation $(x_h, \widetilde{y}_h)$ instead of $(med_{j \in N}\ x_j,\ med_{j \in N}\ \widetilde{y}_j)$. Of course, observation $h \in N$ could be defined as the one such that $x_h = med_{j \in N}\ x_j$.

If $S = \{h \in N \mid x_h \leq med_j\ x_j\}$, $S' = \{h \in N \mid x_h > med_j\ x_j\}$, we obtain

$$DA(X, \widetilde{Y}) = med_{i \in S}\ med_{j \in S'} CWA((x_i, \widetilde{y}_i), (x_j, \widetilde{y}_j)), \tag{10}$$

which amounts to Brown and Mood (1951) technique. A variant of this is taking the set $S$ as the set of the names of the third part of the sample with the smallest $x_j$'s and set $S'$ as the names of the third part of the sample with the highest $x_j$'s, which coincides exactly with the *resistant line method* of Tukey (1970/71). Tukey proposes to choose as the estimate of the slope the value $\widehat{\beta}_1$ such that $med_{i \in S}\ (y_i - \widehat{\beta}_1 x_i) = med_{i \in S'}\ (y_i - \widehat{\beta}_1 x_i)$ and the median of the residuals of both groups $S$ and $S'$ as the estimate of the intercept. It is easy to check that the resistant line method can be written as a CRM estimator.

In what follows, we prove that CRM estimators are strategy-proof in the admissible true sample $\overline{Z}$ such that $\forall i, j \in \{1, ..., n\}$, $x_i \neq x_j$.[6]

[6] We claim that the assumption is no too demanding since variable $x_i$ is real. In case that two different agents $i$ and $j$ share the same $x$, any convention that slightly changes one of them or both for the sake of estimation would work. For example, this one: $\overline{x}_i = x_i - \varepsilon$ if $i = \min\{i, j\}$, for $\varepsilon$ as small as desired. The transformation in the sample is negligible and CRM estimators would be well-defined.

**Theorem 1** *Any CRM estimator is strategy-proof for any admissible true sample $\overline{Z}$.*

**Proof.** Let $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ be the CRM estimator as defined in equations (5), (6) and (7). We prove that for any admissible sample, no agent $i$ behind observation $i$ with true values $(x_i, y_i)$ can ever be better off by reporting a value $\widetilde{y}_i$ different from $y_i$ for any single-peaked preferences $R_i^{y_i} \in \Re^{y_i}$. For simplicity, we shall deal with the case where $\#S'$ and $\#S$ are both odd, but the same proof holds for any other case where $\#S'$ and/or $\#S$ are even with minor changes. We begin by establishing some properties of the CRM estimator. The most efficient way to understand it and avoid large expressions is by means of graphical examples that are nevertheless completely general (see *Figure 2*). Let us think on any sample $(X, y_i, \widetilde{Y}_{-i}) \in \overline{Z}$. For all $x \in E$, we define $\widehat{y}(\widehat{\beta}, x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$. When applying CRM using formulae (5), (6) and (7), notice that by construction, there must always exist two observations, say $k, h \in \{1, ..., n\}$ such that $\widehat{\beta}_1(X, y_i, \widetilde{Y}_{-i}) = \dfrac{\widetilde{y}_h - \widetilde{y}_k}{x_h - x_k}$.[7] Let us call observation $k$ the *directing* observation pointing to observation $h$ (if there are more than one directing observation, take the one with the smallest $x$). Since the admissible domain is $\overline{Z}$, we must distinguish between two possible cases now: either $x_h > x_k$ or $x_k > x_h$ with completely analogous analysis, so we shall focus in just one: $x_h > x_k$. Notice that, since $\widehat{\beta}_1(X, y_i, \widetilde{Y}_{-i}) = \dfrac{\widetilde{y}_h - \widetilde{y}_k}{x_h - x_k}$ and it is the slope corresponding with a median of angles (one for each observation that are themselves medians of angles), there are at least $\dfrac{\#S + 1}{2}$ observations $l \in \{1, ..., n\}$ with

$$med_{\substack{j \in S' \\ j \neq l}} \, CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j)) \geq med_{\substack{j \in S' \\ j \neq k}} \, CWA((x_k, \widetilde{y}_k), (x_j, \widetilde{y}_j)) \qquad (11)$$

and also $\dfrac{\#S - 1}{2}$ observations $l \in \{1, ..., n\}$ would be such that

$$med_{\substack{j \in S' \\ j \neq l}} \, CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j)) < med_{\substack{j \in S' \\ j \neq k}} \, CWA((x_k, \widetilde{y}_k), (x_j, \widetilde{y}_j)). \qquad (12)$$

Now, we can divide the $(x, y)$ plane into four parts taking the directing

[7] Observation $i$ could well be one of those ($k$ or $h$); in that case, we identify the notation $\widetilde{y}_h$ or $\widetilde{y}_k$ with $y_i$. Moreover, notice that $\widehat{\beta}_1(X, y_i, \widetilde{Y}_{-i}) < \infty$ and is always bounded in any domain $\overline{Z}$, since $x_j \neq x_h$ for all $j, h \in N$.

observation $k$ as a reference point as it is illustrated in *Figure 2.1.*: Areas

$$\begin{aligned} A(\widehat{\beta}, X, \widetilde{Y}) &= \left\{(x, y) \in E^2 \text{ such that } x < x_h \text{ and } y > \widehat{y}(\widehat{\beta}, x)\right\}, \\ B(\widehat{\beta}, X, \widetilde{Y}) &= \left\{(x, y) \in E^2 \text{ such that } x > x_h \text{ and } y \geq \widehat{y}(\widehat{\beta}, x)\right\}, \\ C(\widehat{\beta}, X, \widetilde{Y}) &= \left\{(x, y) \in E^2 \text{ such that } x < x_h \text{ and } y \leq \widehat{y}(\widehat{\beta}, x)\right\} \text{ and} \\ D(\widehat{\beta}, X, \widetilde{Y}) &= \left\{(x, y) \in E^2 \text{ such that } x > x_h \text{ and } y < \widehat{y}(\widehat{\beta}, x)\right\}. \end{aligned} \tag{13}$$

Now, notice that since $(x_h, \widetilde{y}_h) \in B(\widehat{\beta}, X, \widetilde{Y})$, by construction of $\widehat{\beta}$, since $n$ is even, area $B(\widehat{\beta}, X, \widetilde{Y})$ contains at least $\dfrac{\#S' + 1}{2}$ sample observations $(x_l, \widetilde{y}_l) \in (X, \widetilde{Y})$. That leaves at most $\dfrac{\#S' - 1}{2}$ sample observations for total area $A(\widehat{\beta}, X, \widetilde{Y}) \cup C(\widehat{\beta}, X, \widetilde{Y}) \cup D(\widehat{\beta}, X, \widetilde{Y})$.

First, we prove that for any observation $(x_l, \widetilde{y}_l) \in B(\widehat{\beta}, X, \widetilde{Y})$, it holds that

$$med_{\substack{j \in S' \\ j \neq l}} CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j)) < med_{\substack{j \in S' \\ j \neq k}} CWA((x_k, \widetilde{y}_k), (x_j, \widetilde{y}_j)). \tag{14}$$

By contradiction, assume that $\exists (x_l, \widetilde{y}_l) \in B(\widehat{\beta}, X, \widetilde{Y})$ such that (14) does not hold; then, it must hold that the area to the right of the line $x = x_l$ and above the straight line that passes through $(x_l, \widetilde{y}_l)$ and the slope defined by the angle $med_{\substack{j \in S' \\ j \neq l}} CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j))$ is always strictly contained in $B(\widehat{\beta}, X, \widetilde{Y})$. By definition of $med_{\substack{j \in S' \\ j \neq l}} CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j))$, there must be at least $\dfrac{\#S' + 1}{2}$ observations in that area, which enters into contradiction with $(x_k, \widetilde{y}_k)$ being the directing observation pointing to $(x_h, \widetilde{y}_h)$, since $(x_h, y_h) \in B(\widehat{\beta}, X, \widetilde{Y})$ but cannot enter into the count for finding $med_{\substack{j \in S' \\ j \neq l}} CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j))$. Now, since all observations $(x_l, \widetilde{y}_l) \in B(\widehat{\beta}, X, \widetilde{Y})$ have a smaller $med_{\substack{j \in S' \\ j \neq l}} CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j))$ than $med_{\substack{j \in S' \\ j \neq k}} CWA((x_k, \widetilde{y}_k), (x_j, \widetilde{y}_j))$, it must be true by (12) that all other observations in $A(\widehat{\beta}, X, \widetilde{Y}) \cup C(\widehat{\beta}, X, \widetilde{Y}) \cup D(\widehat{\beta}, X, \widetilde{Y})$ are such that

$$med_{\substack{j \in S' \\ j \neq l}} CWA((x_l, \widetilde{y}_l), (x_j, \widetilde{y}_j)) \geq med_{\substack{j \in S' \\ j \neq k}} CWA((x_k, \widetilde{y}_k), (x_j, \widetilde{y}_j)). \tag{15}$$

Now, we prove that there is no sample observations in area $A(\widehat{\beta}, X, \widetilde{Y})$ or in other words, $A(\widehat{\beta}, X, \widetilde{Y}) \cap (X, \widetilde{Y}) = \emptyset$. By contradiction, suppose that there is one, say $(x_l, \widetilde{y}_l) \in A(\widehat{\beta}, X, \widetilde{Y})$. By (15), $l$ is such that

$med_{\substack{j\in S'\\ j\neq l}}\ CWA((x_l,\widetilde{y}_l),(x_j,\widetilde{y}_j)) \geq med_{\substack{j\in S'\\ j\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$, so by (12), it is clear that there must exist at least other observation $(x_r,\widetilde{y}_r)$ such that $x_l < x_r < x_k$ and above the straight line passing through $(x_l,\widetilde{y}_l)$ with the slope defined by the angle $med_{\substack{j\in S'\\ j\neq l}}\ CWA((x_l,\widetilde{y}_l),(x_j,\widetilde{y}_j))$. Now, we focus in observation $r$ in $A(\widehat{\beta},X,\widetilde{Y})$ that maximizes $\left(\widetilde{y}_r - \widehat{\beta}_1 x_r\right)$, i.e., $r\in\{1,...,n\}$ is such that $\widetilde{y}_r - \widehat{\beta}_1 x_r \geq \widetilde{y}_j - \widehat{\beta}_1 x_j,\ \forall j\in A(\widehat{\beta},X,\widetilde{Y})$. By construction, there cannot be any other observation in $A(\widehat{\beta},X,\widetilde{Y})$ included into $r$'s "counting space", i.e., the space to the right of $x_r$ and above $med_{j\neq r}\ CWA((x_r,\widetilde{y}_r),(x_j,\widetilde{y}_j)) \geq med_{j\neq k}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$ (see *Figure 2.2.*). Two cases must be analyzed.

**Case 1**:

$med_{\substack{j\in S'\\ j\neq r}}\ CWA((x_r,\widetilde{y}_r),(x_j,\widetilde{y}_j)) > med_{\substack{j\in S'\\ j\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$: by (12), we find a contradiction and the existence of observation $j$ impedes the existence of any observation in $A(\widehat{\beta},X,\widetilde{Y})$.

**Case 2**:

$med_{\substack{j\in S'\\ j\neq r}}\ CWA((x_r,\widetilde{y}_r),(x_j,\widetilde{y}_j)) = med_{\substack{j\in S'\\ j\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$. In this case, there is a tie between two directing observations, i.e., $(x_k,\widetilde{y}_k)$ and $(x_r,\widetilde{y}_r)$, both pointing to two different observations: observation $k$ to $(x_h,\widetilde{y}_h)$ and observation $r$ to, say, $(x_f,\widetilde{y}_f)$. Let us define areas $A'(\widehat{\beta},X,\widetilde{Y})$, $B'(\widehat{\beta},X,\widetilde{Y})$, $C'(\widehat{\beta},X,\widetilde{Y})$ and $D'(\widehat{\beta},X,\widetilde{Y})$ as the analogous to $A(\widehat{\beta},X,\widetilde{Y})$, $B(\widehat{\beta},X,\widetilde{Y})$, $C(\widehat{\beta},X,\widetilde{Y})$ and $D(\widehat{\beta},X,\widetilde{Y})$ in (13) respectively when taking the tied directing observation $(x_r,\widetilde{y}_r)$ as the reference point in the definitions (see *Figure 2.2.*). Now, if $(x_f,\widetilde{y}_f)\in B(\widehat{\beta},X,\widetilde{Y})$ (remember that $B(\widehat{\beta},X,\widetilde{Y})$ is defined as the corresponding region for the reference directing point $(x_k,\widetilde{y}_k)$), the same argument in (12) yields the conclusion that $med_{\substack{j\in S'\\ j\neq r}}\ CWA((x_r,\widetilde{y}_r),(x_j,\widetilde{y}_j)) < med_{\substack{j\in S'\\ h\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$, which we proved to be impossible. If $(x_f,\widetilde{y}_f)\in A(\widehat{\beta},X,\widetilde{Y})$, notice that there cannot be any other observations in area

$A(\widehat{\beta},X,\widetilde{Y})\cap B'(\widehat{\beta},X,\widetilde{Y})$ apart from $(x_f,\widetilde{y}_f)$, by definition of $(x_r,\widetilde{y}_r)$, so area $B(\widehat{\beta},X,\widetilde{Y})\cap B'(\widehat{\beta},X,\widetilde{Y})$ contains $\dfrac{\#S'-1}{2}$ sample observations with strictly smaller median slopes than

$med_{\substack{j\in S'\\ j\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$ and therefore, area $B(\widehat{\beta},X,\widetilde{Y})\backslash B'(\widehat{\beta},X,\widetilde{Y})$

contains only $(x_h, \widetilde{y}_h)$. Now, notice that it must be the case that

$$\begin{aligned} med_{\substack{j\in S' \\ j\neq f}}\, CWA((x_f,\widetilde{y}_f),(x_j,\widetilde{y}_j)) &< med_{\substack{j\in S' \\ j\neq k}}\, CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j)) = \\ &= med_{\substack{j\in S' \\ j\neq r}}\, CWA((x_r,\widetilde{y}_r),(x_j,\widetilde{y}_j)) \end{aligned}$$

and since $f$'s slope is smaller than that of $k$ and $r$ and all $\dfrac{\#S'-1}{2}$ observations in $B(\widehat{\beta}, X, \widetilde{Y}) \cap B'(\widehat{\beta}, X, \widetilde{Y})$ are obviously included in $f$'s own $B$ set, neither $k$ nor $r$ can be the directing observations: $f$ will be the directing one, so we find a contradiction with our assumptions and both cases are covered: the set $A(\widehat{\beta}, X, \widetilde{Y}) \cap (X, \widetilde{Y})$ must be empty and moreover, no possible tie can emerge in the directing observation lying outside the straight line passing through $(x_r, \widetilde{y}_r)$ with slope $\dfrac{\widetilde{y}_h - \widetilde{y}_k}{x_h - x_k}$.

Therefore, we know that all observations $(x_{l,}\widetilde{y}_l)$ in $C(\widehat{\beta}, X, \widetilde{Y})$ and $D(\widehat{\beta}, X, \widetilde{Y})$ are such that $med_{\substack{j\in S' \\ j\neq l}}\, CWA((x_l,\widetilde{y}_l),(x_j,\widetilde{y}_j)) > med_{\substack{j\in S' \\ j\neq k}}\, CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$ and all $(x_{l,}\widetilde{y}_l)$ in $B(\widehat{\beta}, X, \widetilde{Y})$ are such that $med_{\substack{j\in S' \\ j\neq l}}\, CWA((x_l,\widetilde{y}_l),(x_j,\widetilde{y}_j)) < med_{\substack{j\in S' \\ j\neq k}}\, CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$.

The last fact we need in order to prove that CRM estimators are strategy-proof is that it must happen that $\widehat{\beta}_0 = med_{i\in N}\left(\widetilde{y}_i - \widehat{\beta}_1 x_i\right) = y_k - \widehat{\beta}_1 x_k$, that is, the estimate for the intercept term is given by the intercept of the straight line with slope $\dfrac{\widetilde{y}_h - \widetilde{y}_k}{x_h - x_k}$ that passes through the directing observation $k \in \{1, ..., n\}$. This becomes obvious since we know now that area $A(\widehat{\beta}, X, \widetilde{Y})$ is empty and area $B(\widehat{\beta}, X, \widetilde{Y})$ contains $\dfrac{\#S+1}{2}$ sample observations, so the straight line defining the areas $A, B, C$ and $D$ with supports in $(x_k, \widetilde{y}_k)$ and $(x_h, \widetilde{y}_h)$ is in fact the regression line.

Now, we shall prove that no agent $i \in \{1, ..., n\}$ behind any observations can gain by reporting a different $\widetilde{y}_i \neq y_i$. First, let us assume that the sample is $(X, y_i, \widetilde{Y}_{-i})$. Notice that if $i = k$ or any other observation lying on the regression line cannot have any incentive to lie, since the prediction for their $x$'s are exactly their true value (the residual is the smallest possible for any single-peaked preferences he might have). We must only care about the observations outside the regression line. Let us distinguish two cases.

**Case 1**: Let $i \in \{1, ..., n\}$ be such that $x_i < x_k$.

Since $i \in C(\widehat{\beta}, X, \widetilde{Y})$, $y_i$ lies below the regression line and declaring a lower $\widetilde{y}_i < y_i$ increases $med_{\substack{j\in S'\\ j\neq i}}\ CWA((x_i,\widetilde{y}_i),(x_j,\widetilde{y}_j))$,[8] potentially increases $med_{\substack{j\in S'\\ j\neq l}}\ CWA((x_l,\widetilde{y}_l),(x_j,\widetilde{y}_j))$ for $l \in S$ with $(x_l,\widetilde{y}_l) \in B(\widehat{\beta},X,\widetilde{Y})$ (it cannot decrease them in any case), but it cannot make them be higher than $med_{\substack{j\in S'\\ j\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$ and has no effect on the slopes $med_{\substack{j\in S'\\ j\neq l}}\ CWA((x_l,\widetilde{y}_l),(x_j,\widetilde{y}_j))$ for $l \in S$ with $(x_l,\widetilde{y}_l) \in B(\widehat{\beta},X,\widetilde{Y})$ and $l = k$. The net effect is that the median of the new slopes for all observations cannot change $\widehat{\beta}_1$ and therefore cannot lower $i$'s prediction, so there is no incentive to report a smaller value. Now, imagine that agent $i$ declares a higher $\widetilde{y}_i > y_i$. If $\widetilde{y}_i < \widehat{\beta}_1(X,y_i,\widetilde{Y}_{-i})x_i$, $i$'s slope decreases, but it cannot be smaller than $med_{\substack{j\in S'\\ j\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$. Once more it cannot affect the slopes of agents in area $D(\widehat{\beta},X,\widetilde{Y})$ and cannot increase the slopes of the agents in $B(\widehat{\beta},X,\widetilde{Y})$, so the net effect is that observation $k$ continues to be the directing one and there is no gain from lying for the agent behind $i$. If $\widetilde{y}_i > \widehat{\beta}_1(X,y_i,\widetilde{Y}_{-i})x_i$, observation $i$ substitutes observation $k$ as the directing one up to some limit for which a different observation will play the directing role. When $i$ plays the directing observation role, initially it points to observation $k$ and takes the regression line away from the true observation. There is no way such that agent $i$ can reduce its predicted value by increasing $\widetilde{y}_i$ even further, since any other directing observation in $B(\widehat{\beta},X,\widetilde{Y})$ would have a smaller slope than $\widehat{\beta}_1(X,y_i,\widetilde{Y}_{-i})$ and would pass through a point in $B(\widehat{\beta},X,\widetilde{Y})$.

**Case 2**: Let $i \in \{1,...,n\}$ be such that $x_i > x_k$.

If $y_i \in B(\widehat{\beta},X,\widetilde{Y})$ reporting a higher value $\widetilde{y}_i > y_i$ cannot make the slope for $i$ be higher than $\widehat{\beta}_1(X,y_i,\widetilde{Y}_{-i})$, since $med_{\substack{j\in S'\\ j\neq i}}\ CWA((x_i,\widetilde{y}_i),(x_j,\widetilde{y}_j)) < med_{\substack{j\in S'\\ j\neq k}}\ CWA((x_k,\widetilde{y}_k),(x_j,\widetilde{y}_j))$,and the same happens for any observation in $B(\widehat{\beta},X,\widetilde{Y})$. By contrast, the increase in $\widetilde{y}_i$ cannot make the slopes of observations in $C(\widehat{\beta},X,\widetilde{Y})$ and $D(\widehat{\beta},X,\widetilde{Y})$ smaller than $\widehat{\beta}_1(X,y_i,\widetilde{Y}_{-i})$. Since no slope in the sample jumps over the median defining $\widehat{\beta}_1(X,y_i,\widetilde{Y}_{-i})$, the regression line remains unaltered. Now, in order to analyze the possibilities of manipulation when reporting a smaller value $\widetilde{y}_i < y_i$: we must distinguish two cases:

**Case 2.1.**: If $i$'s angle points to some other observation $l$ with $x_l > x_i$, reporting a smaller $\widetilde{y}_i < y_i$ such that $(\widetilde{y}_i,x_i) \in B(\widehat{\beta},X,\widetilde{Y})$ raises the angle

[8] Notice that whenever $i \notin S$, his own angle will not affect to the total count to find the directing angle defining the regression line and therefore $i$ cannot be the directing angle, but he still can affect other's angles and hence the selection of the directing angle.

corresponding to $i$ but cannot reach the even higher angle of the directing observation. Moreover, no other observation in $B(\widehat{\beta}, X, \widetilde{Y})$ surpasses the directing angle due to $i$'s lie. Observations in $C(\widehat{\beta}, X, \widetilde{Y})$ and $D(\widehat{\beta}, X, \widetilde{Y})$ either lower their slopes or do not change, but their slopes cannot jump over the directing angle in any case, so there is no chance of changing the regression line and the predicted $\widehat{y}_i$ until $\widetilde{y}_i$ enters area $D(\widehat{\beta}, X, \widetilde{Y})$. If this is the case, still when $i$'s slope is smaller than that of the directing observation $k$, $i$'s angle decreases to pass through $(\widetilde{y}_i, x_i)$ if $i \in S$, following $i$'s fall in $\widetilde{y}_i$, since no angle of observations in $B(\widehat{\beta}, X, \widetilde{Y})$ can jump on $k$'s new angle (see *Figure 2.3.*). This change lowers the predicted $\widehat{y}_i$, so $i$ will be worse off. If $\widetilde{y}_i$ decreases even more, it will eventually leave the regression line fixed at a lower level, but no other directing point in $B(\widehat{\beta}, X, \widetilde{Y})$, $C(\widehat{\beta}, X, \widetilde{Y})$ or $D(\widehat{\beta}, X, \widetilde{Y})$ will be able to support a new directing angle (see *Figure 2.3.*). Therefore, by lowering $\widehat{y}_i$, the agent behind observation $i$ can only worsen his predicted value.

**Case 2.2.**: If $i$'s angle points to some other observation $l$ with $x_l < x_i$, reporting a smaller $\widetilde{y}_i < y_i$ such that $(\widetilde{y}_i, x_i) \in B(\widehat{\beta}, X, \widetilde{Y})$ lowers $i$'s own angle and cannot make any slope of observations in $B(\widehat{\beta}, X, \widetilde{Y})$ rise over the directing observation one. Analogously, observations in $C(\widehat{\beta}, X, \widetilde{Y})$ and $D(\widehat{\beta}, X, \widetilde{Y})$ can potentially lower their own slopes, but never as low as $i$'s directed angle (see *Figure 2.4.*). Therefore, the regression line does not change. When $\widetilde{y}_i$ falls below the regression line, i.e., when $(\widetilde{y}_i, x_i) \in D(\widehat{\beta}, X, \widetilde{Y})$, by the same reasoning as above, the regression could rotate around the directing observation $(\widetilde{y}_k, x_k)$ pointing to $i$ until eventually it could stop pointing to other one in $D(\widehat{\beta}, X, \widetilde{Y})$, always making the agent behind observation $i$ being worse off for every single-peaked preference that he might have.

Finally, if $y_i \in D(\widehat{\beta}, X, \widetilde{Y})$, lowering $\widetilde{y}_i < y_i$ cannot neither decrease $i$'s slope nor make any slope of observations in $B(\widehat{\beta}, X, \widetilde{Y})$ rise enough to beat $k$ as the directing observation. Nevertheless, it can decrease some slopes of observations in $D(\widehat{\beta}, X, \widetilde{Y})$, but the existence of observation $j$ precludes that any other can beat $k$ in any case, so the regression line remains the same. The last case is that of $y_i \in D(\widehat{\beta}, X, \widetilde{Y})$ when considering lies such that: $\widetilde{y}_i > y_i$. For all $\widetilde{y}_i \in D(\widehat{\beta}, X, \widetilde{Y})$, $i$'s slope falls but never as much as to pick the directing observation $k$'s slope (this would conflict with our previous results about the slopes of observations in each region). Some other observations in $D(\widehat{\beta}, X, \widetilde{Y})$ might increase their own slopes even further. On the other hand, some observations in $B(\widehat{\beta}, X, \widetilde{Y})$ (some $l$ such that $x_l < x_i$) might increase their slopes and some others might see it lowered (some of those $l \in B(\widehat{\beta}, X, \widetilde{Y})$ for which $x_l > x_i$), but there is no chance of beating the median, so the regression does not change. Finally, increasing $\widetilde{y}_i > y_i$ so far such that $\widetilde{y}_i \in B(\widehat{\beta}, X, \widetilde{Y})$ (see *Figure 2.4.*) must make $i$'s slope fall below $k$'s slope, and therefore jumping to the other side of the median. Other observations in areas $C(\widehat{\beta}, X, \widetilde{Y})$ and $D(\widehat{\beta}, X, \widetilde{Y})$ might increase their slopes and some others in $B(\widehat{\beta}, X, \widetilde{Y})$ like $l$ with $x_l < x_i$ might increase their slope, but never as much as to beat $k$ as the directing observation. Some observations in $B(\widehat{\beta}, X, \widetilde{Y})$ like $l$ with $x_l > x_i$ might see their slopes lowered. Finally, the directing observation $k$'s slope must rise initially pointing to $i$ since an observation in area $D(\widehat{\beta}, X, \widetilde{Y})$ has moved up to counting area $B(\widehat{\beta}, X, \widetilde{Y})$. The results of all this is that the regression line moves further away from $y_i$ and for any single-peaked preferences, the agent behind observation $i$ is worse off. Pushing the lie $\widetilde{y}_i$ even further in $B(\widehat{\beta}, X, \widetilde{Y})$ might eventually stop the regression line rotating upwards around $k$ to point to any other observation in $B(\widehat{\beta}, X, \widetilde{Y})$ and no new shift in the regression line emerges. Therefore, we have exhausted all possible cases and there is no chance of getting a better prediction by manipulating the declared response value, so the CRM estimator $\widehat{\beta}$ is strategy-proof. ■

[Insert *Figure 2* about here]

CRM estimators are proved to be strategy-proof for very large admissible samples, since $\overline{Z}$ includes every observation such that $\forall i, j \in N,\ x_i \neq x_j$. Furthermore, CRM estimators possess other interesting statistical properties. When all the observations lie on the same straight line, CRM estimators always capture the line. Moreover, CRM estimators are *scale equivariant* for any true sample $\overline{Z}$. An estimator $\widehat{\beta}$ is scale equivariant if $\forall (X, Y) \in \overline{Z}$, $\forall \lambda \in E$, $\widehat{\beta}(X, \lambda Y) = \lambda \widehat{\beta}(X, Y)$, i.e. a change in the units of measurement of the response variable only re-scales the estimator. Furthermore, CRM estimators also possess good "allocating" properties. In particular, they are all efficient (or Pareto-optimal) in the sense that for all admissible samples and for all individual single-peaked preferences, CRM estimators are such that there does not exist a different regression line that guarantees (weakly) better predictions for all the agents and leaves at least one agent strictly better off. Let us define the efficiency property more rigorously and we shall easily prove the result.

**Definition 4** *Regression estimator* $\widehat{\beta}' = (\widehat{\beta}_0, \widehat{\beta}_1) = T(X, Y)$ *is* Pareto-efficient *if* $\forall Z = (X, Y)$, *there does not exists any* $\beta'_0, \beta'_1 \in E$ *such that*

$\forall i \in N$, $\forall R_i^{y_i} \in \Re^{\widetilde{y}_i}$,

$$[\beta_0' + \beta_1' x_i] \, R_i^{y_i} \left[\widehat{\beta}_0(X,Y) + \widehat{\beta}_1(X,Y)x_i\right]$$

*and* $[\beta_0' + \beta_1' x_j] \, P_j^{y_j} \left[\widehat{\beta}_0(X,Y) + \widehat{\beta}_1(X,Y)x_j\right]$ *for at least one* $j \in N$.

**Proposition 2** *Any CRM estimator is Pareto-efficient for any admissible true sample* $\overline{Z}$.

**Proof.** We know by construction that all CRM estimators are such that for all admissible samples, the regression line always passes through at least two different observations. That means that the prediction offered to at least two individuals is exactly their best-preferred prediction, i.e., their own true value of the response variable. Therefore, at least two individuals are always left as well-off as possible and given the same sample, any other regression line (i.e., $\beta_0', \beta_1' \in E$ ) that either passes only through some of these observations but not through all or does not pass through any of them can only make at least one of them strictly worse-off. Since only one regression line can pass through the set of agents that are given their best predictions, the statement in the definition of Pareto-efficiency above must hold and the proof is complete. ■

Now, we check different well-known estimators in the literature and we shall prove that they are not strategy-proof. It is easy to see that OLS is typically manipulable by any observation. Observations with positive residuals have an incentive to report higher $\widetilde{y}_i$'s and those with negative residuals have an incentive to report smaller values for their response variables.[9] The *least median squares* method (LMS, see Rousseeuw (1984)) is defined as $\widehat{\beta}(LMS) = \arg\max_{\widehat{\beta}} med_{i \in N} \, \widehat{e}_i^2$. For the simple regression case, this method amounts to find the strip containing half of the observations with the smaller width measured in the $y$ axis. The regression line is the straight line that lies exactly in the middle of the strip. Clearly, there are cases where an observation critical for the strip (say, one that lies below the regression line) can lower slightly the reported $\widetilde{y}_i$ and push the regression line closer to his true value. Another median-based estimators for the simple regression model like those of Andrews (1974), Theil (1950), Siegel's repeated median (1982) or Simon's median star (1986) are not strategy-proof either. We briefly explain each of them and will show the possibility of manipulation by means of a simple graphical example. Andrews proposed to classify the observations in

[9] Actually, the only possibility that precludes any gain from manipulating the regression are the cases where all true observations lie on the same straight line.

two sets $L$ and $R$. $L$ would contain those with smaller $x-$value with the exception of a certain fraction of the smallest and $R$ contains those with higher $x-$value with the exception of a certain fraction of those with the highest $x-$value. Moreover, sets $L$ and $R$ do not contain a number of those in a neighborhood of the median $x-$value. Then, Andrews calculates the median of the two remaining subsets of $x-$values, say $med\ x_1$ and $med\ x_2$ and the medians of the corresponding $y-$values of both groups ($med\ y_1$ and $med\ y_2$). The estimation for the slope is then defined as:

$$\widehat{\beta}_1 = \frac{med\ y_2 - med\ y_1}{med\ x_2 - med\ x_1}. \tag{16}$$

*Figure 3.1.* shows a case in which observation $h \in N$ that defines the median of the largest $x-$value group can manipulate the sample by reporting a slightly smaller $\widetilde{y}_h$ (the direction of the manipulation is shown with the straight arrow and the manipulated new regression line is depicted as the bold dashed line). Another well-known robust estimator that can be easily proved not to be strategy-proof is the one proposed by Theil (1950). Theil's estimator for the slope is:

$$\widehat{\beta}_1 = med_{1 \le i < j \le n} \frac{y_j - y_i}{x_j - x_i}. \tag{17}$$

Theil's estimator breakdown point is about 29.3%, but it is not strategy-proof. *Figure 3.2.* provides an example in which the agent behind observation $h \in N$ can profitably manipulate the sample by reporting a higher $\widetilde{y}_h$. An interesting variant of Theil's estimator is the repeated median estimator proposed by Siegel (1982), which has a breakdown point of 50%. The method consists in computing a two-stage median of the pairwise slopes. The estimates for both parameters are:

$$\begin{aligned} \widehat{\beta}_1 &= med_{i \in N}\ med_{j \in N \setminus \{i\}} \left( \frac{y_j - y_i}{x_j - x_i} \right) \\ \widehat{\beta}_0 &= med_{i \in N} \left( y_i - \widehat{\beta}_1 x_i \right). \end{aligned} \tag{18}$$

Siegel's repeated median estimator is not strategy-proof, as *Figure 3.3.* shows. Agent $h \in N$ can obtain a better result by reporting a higher $\widetilde{y}_h$. Notice that there exists an analogous repeated median version among the CRM estimators that is strategy-proof. The difference between the two are clear: while Siegel's repeated median estimator finds the two-stage median of the pairwise slopes of the observations, our CRM estimator obtains the two-stage median of the pairwise clockwise *angles* defined by the slopes. This quite simple change happens to be crucial regarding the strategic properties of both methods. Finally, we consider Simon's (1986) median star estimator, with slope defined as:

$$\widehat{\beta}_1 = med_{i \in N} \left( \frac{y_i - med_{j \in N}\ y_j}{x_i - med_{j \in N}\ x_j} \right). \tag{19}$$

The median star estimator was actually first proposed by Hampel (1975) and amounts to the line passing through $(med_{j \in N}\ x_j,\ med_{j \in N}\ y_j)$ which has equal number of positive residuals on both sides of $med_{j \in N}\ x_j$. *Figure 3.4.* shows a simple example in which agent $h \in N$ can gain by manipulating the sample. Notice that the median star estimator also has a CRM version defined on the directing angles rather than on the slopes themselves, with the additional change of the regression line passing through $(x_i,\ y_i)$ such that $x_i = med_{j \in N}\ x_j$.

[Insert *Figure 3* about here]

# 3 The simulation results

In this section we show some examples and Monte Carlo experiments to motivate the data fits obtained using different linear regression methods. In particular we compare the OLS estimates under some kind of manipulation with three types of CRM estimators, all of them defined in *Section 2*: the simple clockwise version of the "Repeated Median" estimator (i.e., the one defined by equations (5), (8) and (7)), the version of the so called "Resistant Line" method due to Brown and Mood (1951) (equations (5), (10) and (7)) and the clockwise version of the median star estimator (equations (5), (9) and (7)). The samples are simulated from the following data generating processes (DGP hereafter):

$$DGP1: y_i = 5 - 0.5x_i + e_i \text{ where } \quad e_i \sim N(0,1) \tag{20}$$

$$DGP2: y_i = -5 + 0.5x_i + e_i \text{ where } \ e_i \sim N(0,1) \tag{21}$$

Therefore we show the performance of the estimators when fitting either a negative or a positive slope. The unit variance was also used to consider enough sample variation to capture the differences of the data fits but preserving the linear relationship among the variables. Given a particular estimation method, there is scope for different ways of manipulating the sample.

If we are interested in predicting the likely manipulating behavior of agents from a strategic point of view when OLS is applied, we are in trouble, since no pure strategies Nash equilibrium of the strategic OLS estimation induced game exists whenever all the observations do not lie on the same straight line.

Hence, for any declared values of the others, there will be at least one agent interested in declaring a different value. This is so because OLS parameters depend continuously (and with no bounds) on the declared response variable values of the agents and there is always a value of the dependent variable for every agent that can move the regression line until passing exactly through the true value, so this will be each agent's best response. There is no easy way of predicting a stable equilibrium behavior of the agents under OLS estimation and it could well be the case that agents are not perfectly informed either about the others' true values or their potential lies, so we take a shortcut: we assume bounded rationality strategic behavior for the agents by considering that the "magnitude of the lie" is related to the value of the residual, that seems to be a reasonable approximation to the agent's behavior. Furthermore, we limit the manipulation behavior only to those agents whose residuals are bigger (in absolute value) than the variance of the regression, that is, those at risk of being treated as potential true outliers. To simplify, we assume that the agents that obtain predictions "not too far away" from their respective true values will not lie (the gains will not be important enough as to obtain additional information to play strategically).

Therefore, among the agents that will lie, the agents whose residuals are positive (negative) would tend to increase (decrease) their reported value in an attempt to bias the regression towards their true value. The contaminated sample is generated by the following procedure:

$$\widetilde{y}_i = \begin{cases} y_i + c_1 k_i^2 \widehat{e}_i & \text{if } \; k_i \geq c_2 \\ y_i & \text{if } \; k_i < c_2 \end{cases} \tag{22}$$

where $k_i = \dfrac{|\widehat{e}_i|}{\sqrt{\dfrac{1}{n}\sum_{j=1}^{n} \widehat{e}_j^2}}$ , $\widehat{e}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \;\; \forall i = 1, 2, ..., n$ and $\widehat{\beta}_0$ and $\widehat{\beta}_1$ being the OLS estimates obtained with the true sample. Two arbitrary constants, $c_1$ and $c_2$ must be also considered, in particular we used $c_1 = 10$ and $c_2 = 1$ which implies that less than 1/3 of the observations are contaminated on average.

*Figures 4* and *5* show some examples of the estimates obtained using 20 observations simulated from the DGP 1 (20) and DGP 2 (21) respectively.

It is clear from these pictures that some of the fits provided by the estimators within the "Clockwise Repeated Median family" are better than the ones obtained by the contaminated OLS. In particular the "Resistant Line" predictions are pretty close to those of the OLS and both the clockwise "Repeated Median" and "Median Star" seem to be biased somehow[10].

[Insert *Figures 4* and *5* about here]

Finally, *Figures 6* and *7* display the empirical histograms of both the regression intercept and the slope[11]. The histograms have been plotted after simulating 1000 different samples from DGP 1 (20) and 2 (21) and computing the corresponding estimates for both parameters by using the five methods considered in this section. We only show the Figures for DGP 1 since the results of DGP 2 are similar. Moreover the variance of the simulated models was reduced to make the comparisons clearer in the pictures. These Figures show clear evidence supporting the CRM strategy-proof estimates (in particular the Resistant Line) over OLS estimates when the agents could be contaminating strategically the sample. Observe that the Resistant Line estimates are consistent, although with a slower convergence rate than the OLS ones. The other estimates within the CRM family are not so well asymptotically behaved (for example, the clockwise version of the Repeated Median seem to be biased) but the contaminated OLS does not have any desirable property.

[Insert *Figures 6* and *7* about here]

Finally, *Tables 1* and *2* report the mean squared error ratios of the Contaminated OLS (COLS) and the Resistant Line (RL) to the OLS estimates for the regression slope. In order to highlight the merits of the strategy-proof estimators we chose the Resistant Line among all the family and for the sake of brevity we only report the estimates for the regression slope. The estimates were computed from simulations of bivariate normal distributions[12] for

[10] Observe that the further the two observations which select the CRM estimated line are, the better fit is obtained. Therefore the "Resistant Line" seems to be more appropiate since both observations are obtained from two different subsamples.

[11] All the simulations were programmed and run using TSP (Time Series Processor) software.

[12] Note that the distribution of $Y$ conditioned on $X$ when $\rho = 0.928$ and $\rho = -0.928$ corresponds to the DGP's shown in equations (20) and (21) respectively. Moreover we generated a hundred simulations of $n$ observations, but the results do not significantly differ from those obtained when simulating a thousand times instead.

different values of both the correlation coefficient ($\rho$) and the observations ($n$). These tables show the efficiency loss of the estimators compared to the OLS one. In particular RL is a robust and non-manipulable estimator at the cost of duplicating or even triplicating the OLS variance. Nevertheless, if the agents reporting the data behave strategically this is a minimum cost because under the manipulation considered, the contaminated OLS mean squared errors could be bigger than three hundred times the true OLS counterparts.

[Insert *Tables 1* and *2* about here, called *Figures 8* and *9* below]

# 4 Conclusions

In this paper we have introduced a class of estimators for the simple regression case that are immune to strategic contamination of the data when the agents behind the observations are interested in not being true outliers. CRM estimators are thus recommendable when the lack of information about the response variable values is an important problem. We claim that when information is not easily observable and verifiable, the individual incentives to report false information must be taken into account. Traditionally, statisticians have perhaps underestimated this fact in some cases, where the information-extraction aim of statistics is fundamentally linked with clear and unavoidable economic resource allocation consequences of the information obtained. In this case, a (partial) conflict of interest between the researcher and the agents that provide the data must be solved and the way in which information is aggregated (the estimators or regression method used) can potentially change the incentives of the individuals to declare their true values. There is a gain in credibly committing to using a specific way of aggregating information. If the agents who have the private information perceive that their own well-being depend on their reported information, they will try to act strategically and the sample will be contaminated, but if they believe that the econometrician is going to use a strategy-proof estimator that leaves no gain in reporting false information, the agents behind the observations will behave truthfully, the data will be trustworthy and the information reliable.

Our approach to strategic data contamination is just a particular case appropriate when the information extracted from the regression is likely to affect the interests of the individuals that report the information about the response variable in a particular direction. However, there may be different social situations where the estimates are used (or perceived to be used) in a different way and might change the individual incentives from those analyzed in this paper. For example, it is possible to imagine contexts where the agents behind the observations could be interested in reducing, say, the difference between their true $y_i$ value and the average of the predicted values, $\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i$, or not being perceived as an outlier regarding the reported value, i.e., minimizing $|\widehat{y}_i - \widetilde{y}_i|$ instead of $|\widehat{y}_i - y_i|$. All these cases can nevertheless be analyzed within our strategic approach, but changing the preferences the agents have. Strategy-proofness still holds as a strong incentive compatibility requirement, but individual preferences that summarize their incentives and guide their strategic behavior when facing any estimator will no longer be single-peaked. Further research on this topic will presumably lead to very different estimators.

Notice, however, that different strategy-proof CRM estimators also perform very differently in terms other than being resistant to some specific kind of data manipulation in large domains, although they are calculated in a similar way. In particular, the resistant line, either discarding some observations or not, is a particularly interesting method that seem to have high consistency and robustness, while other CRM estimators like the clockwise repeated median or the clockwise median star do not provide very good fits. Nevertheless, if the problem of strategic data contamination is important enough, the loss of consistency involved in CRM estimators is a small price to pay in exchange for the accuracy of the reported data implied by using strategy-proof estimators. The exact nature of the trade-off between estimation consistency and incentive compatibility is nevertheless still unclear and further research on the issue could be worthwhile.

# References

[1] Andrews, D. F., 1974. A Robust Method for Multiple Linear Regression, Technometrics, 16, 523-531.

[2] Barberà, S., Jackson, M., 1994. A characterization of Strategy- Proof Social Choice Functions for Economies with Pure Public Goods. Social Choice and Welfare 11, 241-252.

[3] Black, D., 1958. The Theory of Committees and Elections, Cambridge University Press, London.

[4] Brown, G. W., Mood, A. M., 1951. On Median Tests for Linear Hypothesis, in Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability, edited by J. Neyman, University of California Press, Berkeley and Los Angeles, 159-166.

[5] Gibbard, A., 1973. Manipulation of Voting Schemes: A General Result. Econometrica 41, 587-601.

[6] Hampel, F. R., 1975. Beyond Location Parameters: Robust Concepts and Methods, Bull. Int. Stat. Inst., 46, 375-382.

[7] Moulin, H., 1980. On Strategy- proofness and Single- peakedness. Public Choice 35, 437-455.

[8] Rousseeuw, P.J. and Leroy, A.M., 1987. Robust Regression and Outlier Detection, John Wiley & Sons, Inc., New York.

[9] Rousseeuw, P.J., 1984. Least Median Squares Regression, Journal of the American Statistical Association, 79, 871-880.

[10] Siegel, A. F., 1982. Robust Regression Using Repeated Medians. Biometrika, 69, 242-244.

[11] Simon, S. D., 1986. The Median Star: an Alternative to the Tukey Resistant Line, paper presented at the Joint Statistical Meetings, Chicago, 18-21 August.

[12] Theil, H., 1950. A Rank-invariant Method of Linear and Polynomial Regression Analysis (Parts 1-3), Ned. Akad. Wetensch. Proc. Ser. A, 53, 386-392, 521-525, 1397-1412.

[13] Tukey, J. W., 1970/71. Exploratory Data Analysis (Limited Preliminary Edition), Adison-Wesley, Reading, MA.

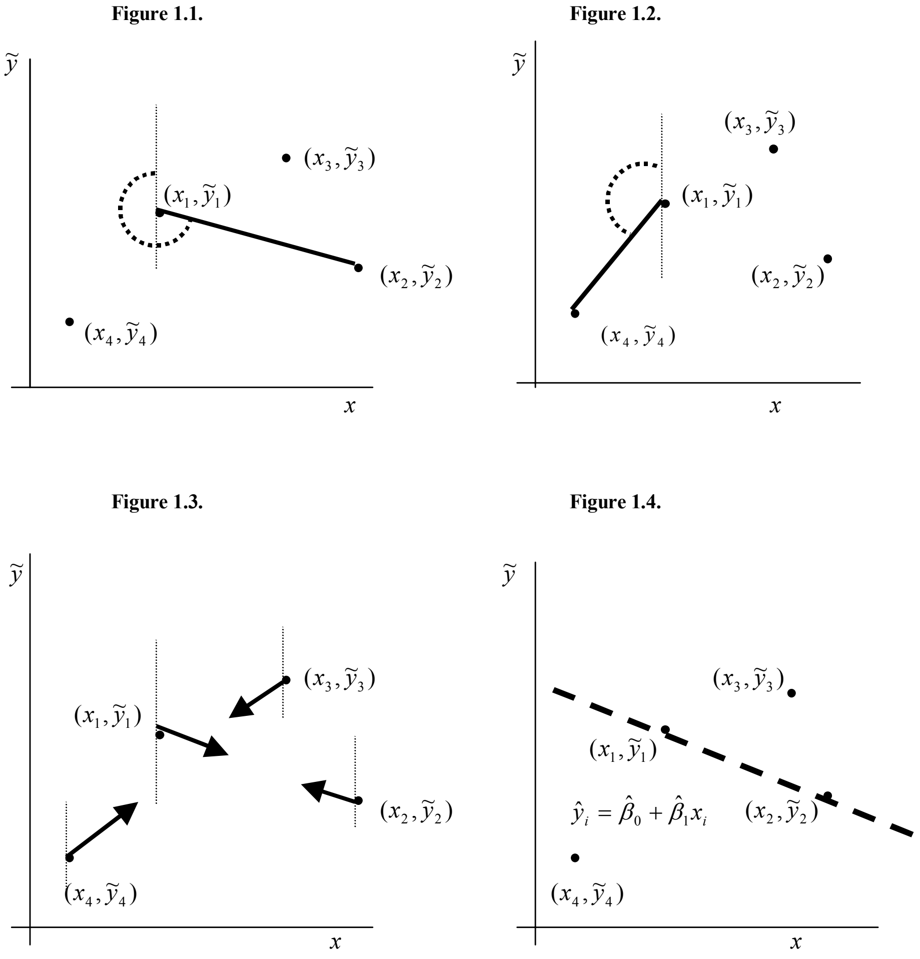


Figure 1.1.

Figure 1.2.

Figure 1.3.

Figure 1.4.

Figure 1:

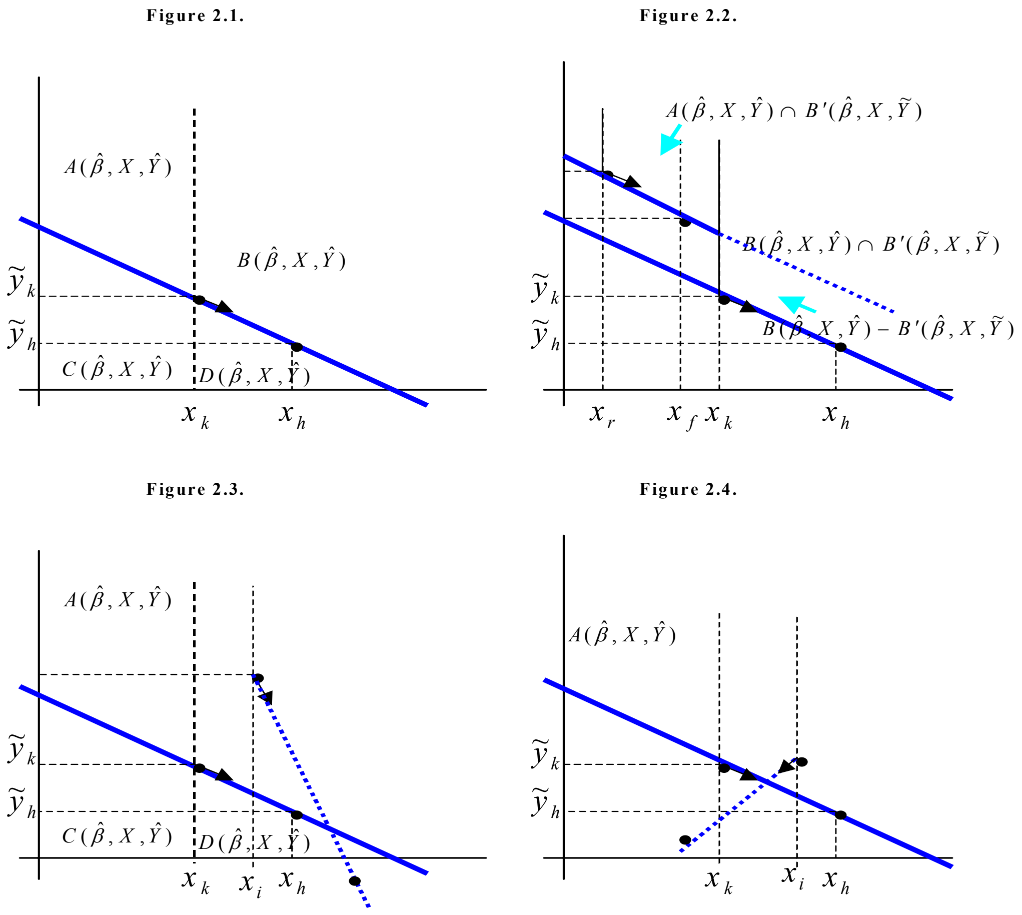


Figure 2.1.

Figure 2.2.

Figure 2.3.

Figure 2.4.

$A(\hat{\beta}, X, \hat{Y})$
$B(\hat{\beta}, X, \hat{Y})$
$C(\hat{\beta}, X, \hat{Y})$
$D(\hat{\beta}, X, \hat{Y})$
$\tilde{y}_k$
$\tilde{y}_h$
$x_k$
$x_h$
$A(\hat{\beta}, X, \hat{Y}) \cap B'(\hat{\beta}, X, \tilde{Y})$
$B(\hat{\beta}, X, \hat{Y}) \cap B'(\hat{\beta}, X, \tilde{Y})$
$B(\hat{\beta}, X, \hat{Y}) - B'(\hat{\beta}, X, \tilde{Y})$
$x_r$
$x_f$
$x_i$

Figure 2:

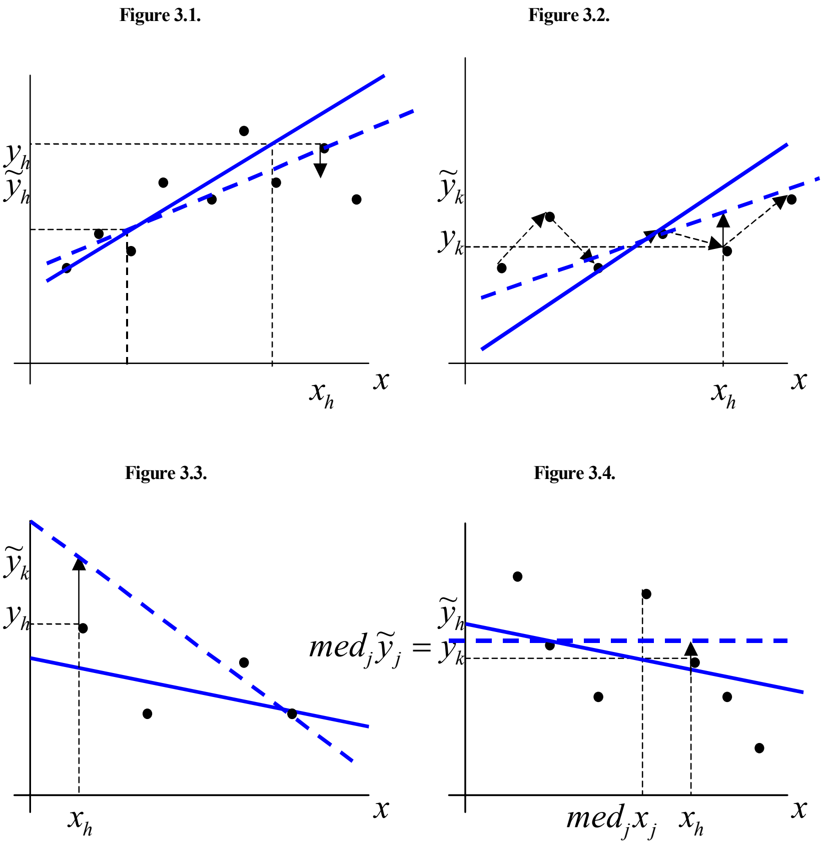


Figure 3.1.

Figure 3.2.

Figure 3.3.

Figure 3.4.

Figure 3:

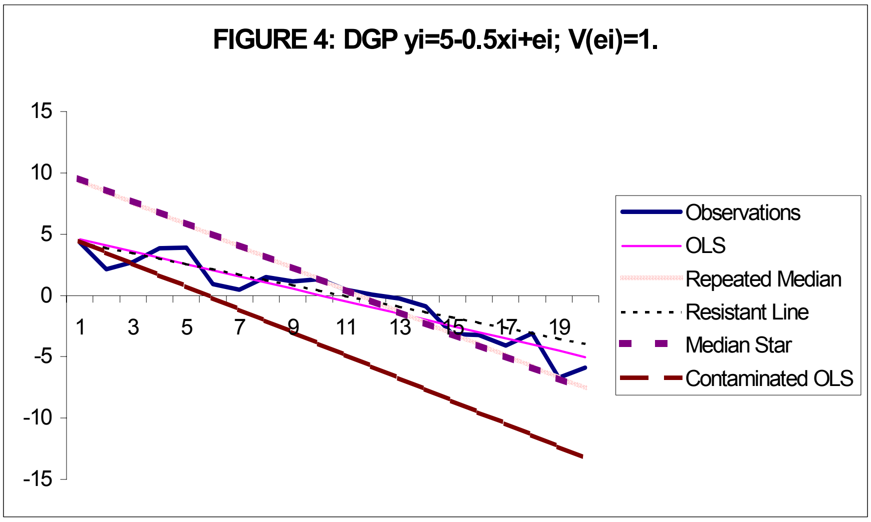


Figure 4:

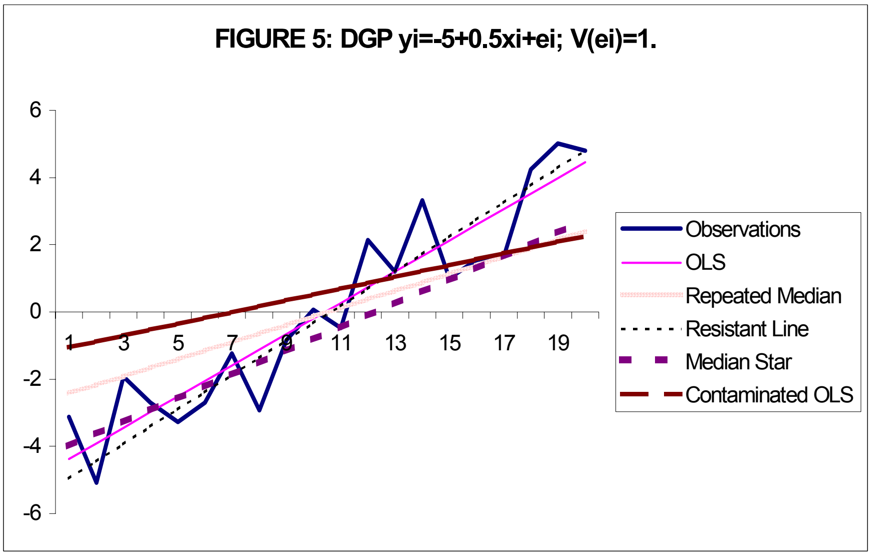


Figure 5:

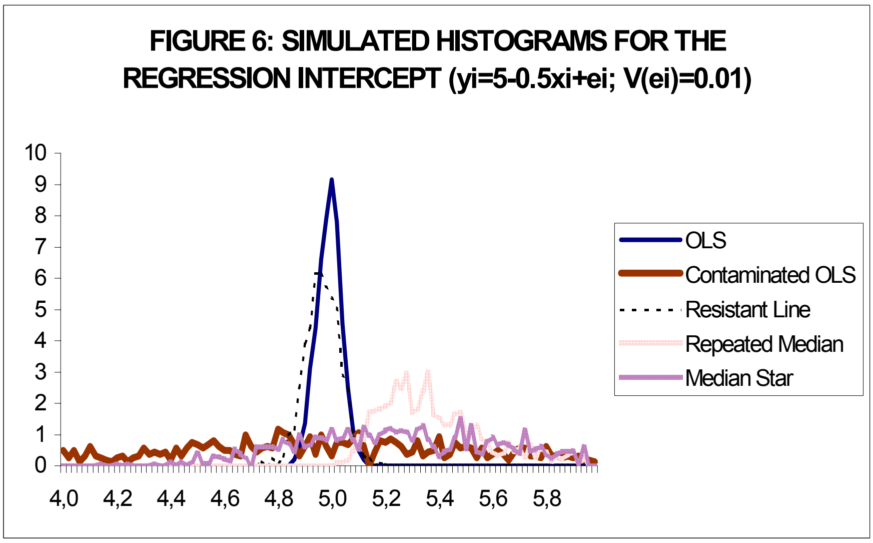


Figure 6:

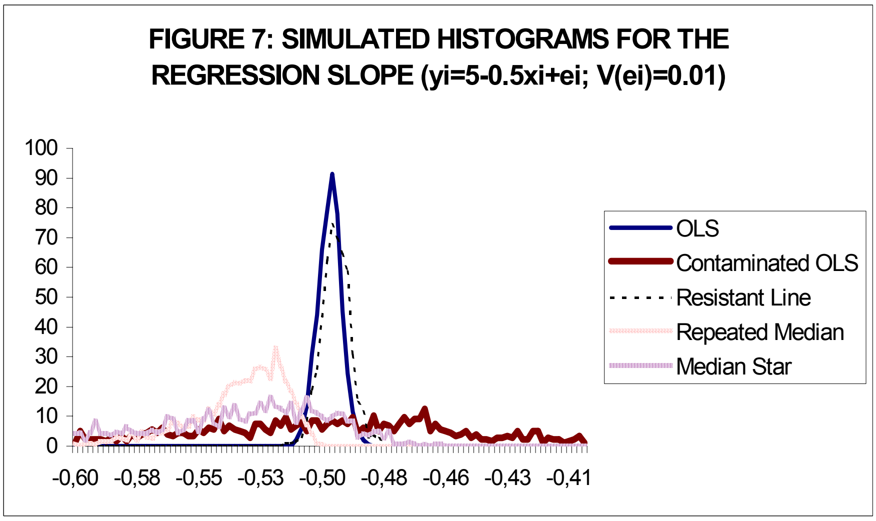
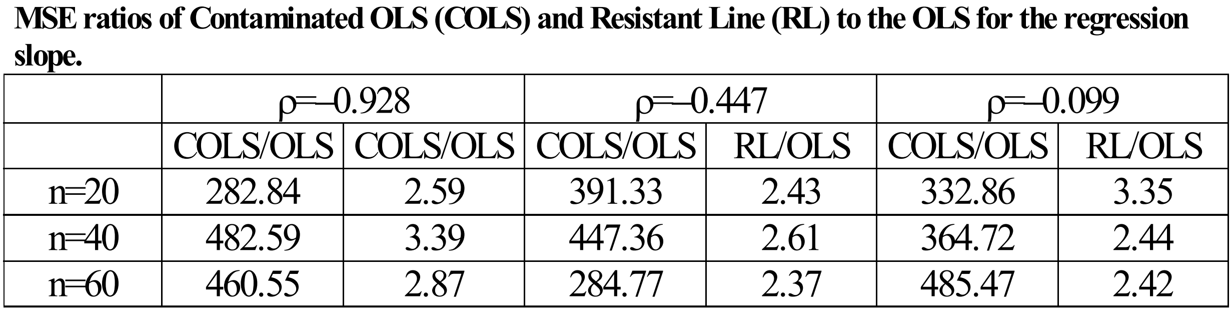


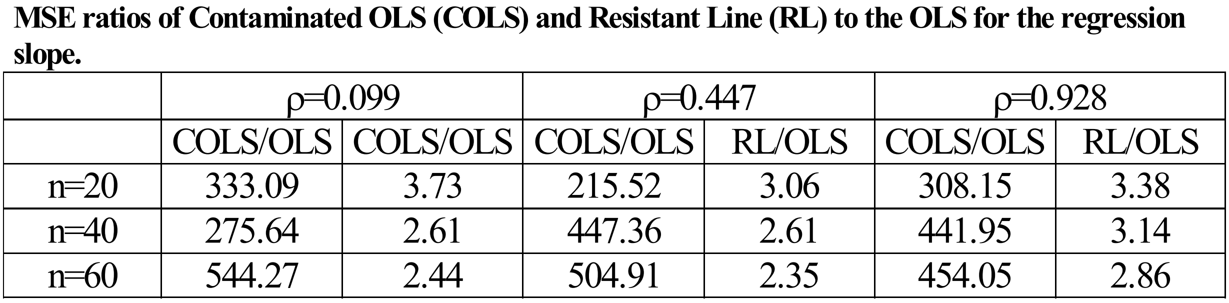
Figure 7:



**MSE ratios of Contaminated OLS (COLS) and Resistant Line (RL) to the OLS for the regression slope.**

| | ρ=–0.928 | | ρ=–0.447 | | ρ=–0.099 | |
|---|---|---|---|---|---|---|
| | COLS/OLS | COLS/OLS | COLS/OLS | RL/OLS | COLS/OLS | RL/OLS |
| n=20 | 282.84 | 2.59 | 391.33 | 2.43 | 332.86 | 3.35 |
| n=40 | 482.59 | 3.39 | 447.36 | 2.61 | 364.72 | 2.44 |
| n=60 | 460.55 | 2.87 | 284.77 | 2.37 | 485.47 | 2.42 |

Figure 8:



**MSE ratios of Contaminated OLS (COLS) and Resistant Line (RL) to the OLS for the regression slope.**

| | ρ=0.099 | | ρ=0.447 | | ρ=0.928 | |
|---|---|---|---|---|---|---|
| | COLS/OLS | COLS/OLS | COLS/OLS | RL/OLS | COLS/OLS | RL/OLS |
| n=20 | 333.09 | 3.73 | 215.52 | 3.06 | 308.15 | 3.38 |
| n=40 | 275.64 | 2.61 | 447.36 | 2.61 | 441.95 | 3.14 |
| n=60 | 544.27 | 2.44 | 504.91 | 2.35 | 454.05 | 2.86 |

Figure 9: