Exploring the Proximate Cause of MNC Hegemony

Swapan Kumar Chakrabarti
Reader, Department of Commerce,
University of Calcutta

Abstract

Since the beginning of the process of globalization in the early 1990s, the Multinational Corporations (MNCs) have figured crucially well to constitute a major vehicle of the process. These giant international producers are the principal agents of Foreign Direct Investments in different host nations. The objective of this paper is to pinpoint the precise manner in which the MNCs, despite their several disadvantages in the host countries, dominate the domestic firms of the latter. The analysis shows that they are able to do so because of their substantial cost advantage over their host country counterparts.

Key-Words : Hegemony ; Multinational ; Corporation ; Foreign Direct Investment ; Knowledge-based Assets ; Representative rival unit.

After the end of the cold war, a large section of the world economies has been swayed by a philosophy of internal and external economic liberalization as part of their reform measures, the hallmark of which is a free and fast flow of technology, investment, output and information across the nations. The phenomenon is stylized in the term 'globalization' that has, since the early 1990s, become a common word, perhaps even to the layman. Of significance, the Multinational Corporations (MNCs) have come to the limelight in the process, their activities having picked up with unprecedented buoyancy. The World Investment Reports (1997, 1998) indicate that nearly one-fifth of the world GNP is produced by the MNCs alone. An MNC, it is well-known, is a giant size international producer which being headquartered in a home country invests and produces in a multitude of nations and hence its name. As an MNC steps in onto a particular host country to make real investment there for joining production in a certain existing industry or launching production along a new activity in the latter, it is observed to be sufferering from a number of special types of disadvantages that are in no way experienced by the local firms of the host nations. But what is of wonder is that, despite a whole host of disadvantages, when it comes to the question of competitive performance on an average, the domestic firms in a host country stand nowhere near an MNC, be it in respect of price or product quality. Wherein lies the key to this hegemony? It is precisely this question that this paper addresses itself to. A common argument that is often put forward as an explanation of this phenomenon is superiority of the technology used by the MNCs. Some writers venture to seek the reason in the relatively greater degree of competitive strength of those giant firms. But these arguments are apparently very general in nature and lack in rigour. The objective of this paper is to offer exposition of a simple mathematical model that is capable of providing a formal basis to generate precision of presenting a palpably convincing logic in this behalf.

MNC : The Principal Agent of FDI

When a foreigner invests in a host country to produce and sell a product there, the investment is said to be a foreign direct investment (FDI) of the latter flowing from the concerned foreign country. Theoretically, this investment may be forthcoming from any kind of foreign unit

whatsoever, for example, a foreign individual or a foreign corporate body. However, for all practical purposes, it is chiefly a large foreign corporate entity capable of operating on a global scale, namely, an MNC that causes an FDI in a host country. Note that the investor's name is not associated with description of the FDI. What matters in this regard is the name of the country that the investor belongs to. Thus when an MNC, based in country *A* (home country), undertakes a real investment in country *B* (host country) with an express intention of producing and selling a product in the latter, this investment is said to be an FDI of *B* flowing from *A*. However, the basic point that we are emphatic in laying stress upon must not be missed. The chief message is that it is an MNC that is eventually the principal agent of FDI in any host country. Furthermore, it is FDI that has become the catchword of the ongoing process of globalization.

Challenges Faced by an MNC

In the case of wholly-owned subsidiaries, the multinational enterprises have both ownership and managerial control in full measure over their investment in a host country. In joint venture, of course, the degree of control is obviously less. But, in general, the MNCs encounter a common set of difficulties in the operation of their production centres abroad. Their difficulties abroad are, in fact, legion. However, several distinct instances can be cited. First of all, there are risks on the political front. Significant shifts in the policies and attitudes of the host country towards the MNCs may occur, for the worse to the latter, in the wake of possible change of government. Secondly, there are strains owing to differences in language and culture between the home and the host countries. In a bid to ensure a powerful appeal so as to win hearts of the host country's consumer folk at large and to attract them by creating their genuine credibility and confidence in the product, the multinationals incur massive advertisement expenditure in the local languages of the host country at a much bigger scale vis-a-vis the domestic firms. The question of additional cost is equally applicable for creating an effective marketing infrastructure. Thus the social and cultural differences entail an extra amount of endeavour on the part of the MNCs to sell their product and hence an extra burden of costs for them. In contrast, no special or additional burden as such is required to be borne by the local firms of the host country. Thirdly, for maintaining and operating the production centres in the various host countries, an MNC is necessarily involved in direct costs of transportation including air travel expenses of its executives as required for regular visits of production sites. Added to this, there are costs of communication with the branches and the plants all over the globe in the form of telephone calls, e-mails and the like. Fourthly, the multinationals do not have a sound practical knowledge of the host country's business surroundings, nor are they reasonably well conversant with the laws of taxation in different countries and details of the *modus operandi* of dealing with the government procedures, at least in the initial stages of operations. There is no denying that they often tend to be discriminated against in the host countries so far as tax obligations are concerned. Fifthly, there are risks of exchange rate fluctuations that pose a serious threat on their repatriation of profits. This kind of risks and uncertainties faced by the MNCs does act to constitute a genuine cost for them. This problem is altogether absent in the case of the host country firms. Finally, more often than not, the MNCs have a predilection for putting people from their own home country as top brass level managers and technicians of their branches / plants in the host nations and in order to induce these people to be so posted, they need to offer an attractive and

hence a markedly higher package of wages. Each of these challenges has a cost implication, which, when added up, constitutes a significant package. The only redeeming point, however, is that this challenge cost for an MNC plausibly diminishes over time as an MNC increasingly gets acclimatized to the host countries.

Advantages of the MNCs

Coming to the other side of the story, one may note three stupendous advantages of an MNC, namely, an *ownership advantage*, a *location advantage* and an *internalization advantage*, as suggested by J. Dunning in what has come to be known as 'the OLI framework' (formed by considering the first letter of each type). Any aspect of business that confers an MNC a substantially high share in the market is called the ownership advantage, for example, a production formula that is beyond exact emulation by any rival producer. Location advantage is one by virtue of which the MNC finds it more profitable to produce abroad than to export by producing in the home country. Examples are found in cheap factor cost and low-cost transport in foreign countries or avoidance of export bottlenecks caused by heavy import duties imposed by different nations. Finally, the internalization advantage of a firm refers to its perception to the effect that it can itself exploit the production process internally better than anyone else. It is as a result of this perception that the firm is expected to show a stubborn resistance to any lucrative proposal to sell off its investment to the market.

Knowledge-based Assets

The aforesaid three advantages have a strong potential to outweigh the various challenges faced by an MNC in continuing its overseas production. As a matter of fact, given the latter two advantages, the ownership advantage enables an MNC to gain an edge over its rivals — the domestic firms of the host country. How is it ensured? For the purpose, it is important to have a closer scrutiny of the ownership advantage.

The ownership advantage, as stated earlier, emanates from the possession of certain assets which others do not have any access to. For example, the asset may be some unique technical formula used in production, which gives the product such a rare attribute as a to attract a very large section of the consumers. Being based on the producer's knowledge as they are, these assets are called *Knowledge-based Assets* (KBA). The knowledge-based assets are not necessarily constituted by a patent or an exclusive technical knowledge alone ; these may also include human capital of an exceptional standard like a high quality team of engineers and management people.

Crucial Properties of KBA

The knowledge-based assets of a firm possess a crucially important property. These assets are somewhat like 'public goods' within the different plants of the same firm. The reason is obvious. As a public good can be consumed jointly by many people, so can the knowledge-based assets of a firm be utilized by all the plants of the same firm. Once a firm creates a knowledge-based asset by making certain amount of one-time fixed investment, it can use the same asset in all of its plants without being required to replicate the cost each time for each plant. It is in this sense that these assets are said to have a public good character. Incidentally, there is no reason to suppose that the knowledge-based assets can be created easily. On the contrary, these assets

may be even costlier than all the physical assets used in a particular plant of the firm, taken together. However, one difference between the knowledge-based assets and the physical assets of a firm is very clear. Physical assets are plant-specific assets. If a firm arranges for a physical asset in plant X by way of withdrawing it from plant Y, output in plant Y will obviously suffer. Hence, on every occasion a physical asset is created in a particular plant of a firm, there is the question of incurring a fresh cost. But this is not the case with the knowledge-based assets as explained already.

The MNC Edge : A Model

Now we shall see, on the basis of a simple model, how an MNC attains a competitive edge over its rival firms in the host countries. For the sake of an easy exposition of the model, the following assumptions are made.

(1) An MNC, based in a certain country *A*, has representation in n number of host countries ;

(2) In each host country, the MNC has only one wholly-owned subsidiary unit with only one production centre ;

(3) Every production centre of the MNC is of identical size with identical productive capacity ;

(4) In all countries labour is homogenous and identically productive ;

(5) There is only one rival firm of the MNC in every host country ;

(6) There is only one plant of the rival firm in every host country and let us call it a representative rival unit of the j^{th} host country $(\forall\, j = 1, 2, \ldots\ldots\ldots n)$;

(7) The representative rival unit of the j^{th} host country is of identical size with identical productive capacity and is also identical to every production centre of the MNC, in respect of size and productive capacity $(\forall\, j = 1, 2, \ldots\ldots\ldots n)$;

(8) Every production centre of the MNC as also every representative rival unit of the j^{th} host country produces the same product, say Z, of identical standard ;

(9) Demand conditions and the conditions of variable input costs are the same everywhere ;

(10) Production function is the same everywhere and because of identical demand and cost conditions, the output levels of Z are also the same everywhere ;

(11) The real value of the plant-specific asset of every production centre (in terms of a common numeraire, say labour) is denoted by α and the real value of the knowledge-based assets defined in the same fashion is denoted by β ;

(12) The imputed real value (in terms of labour) of the cost of the challenges faced by the MNC per production centre in a host country is the same and denoted by λ. This is treated as a parameter but this parameter is assumed to decrease asymptotically to a minor figure with passage of time.

It may be observed that the total variable cost in every production centre of the MNC and that of the representative rival unit are the same. As such, we would ignore it and concentrate only on the aspect of fixed cost to scrutinize whether there is any difference between them on this score. Now, by virture of the aforesaid assumptions, the fixed cost in a representative rival unit (RRU) consists of two components : a plant-specific asset (α) and a knowledge-based asset (β) only. Hence the total fixed cost for the RRU is $\alpha+\beta$ ..... (1)

Again, the total fixed cost of the MNC for operating the production centres in n number of countries is obviously $n\lambda+n\alpha+\beta$.

Hence, the total fixed cost per production centre of an MNC is $\lambda+\alpha+\frac{\beta}{n}$.

Now, as the number of host countries tends to be larger, we have

$$\lim_{n\to\infty}\left[\lambda+\alpha+\frac{\beta}{n}\right]=\lambda+\alpha \quad \ldots\ldots (2).$$

Obviously, the MNC unit gains an edge over the RRU only if $\lambda+\alpha<\alpha+\beta$ [from (1) and (2)]

i.e. if $\lambda<\beta$ ..... (3).

Relation (3) may be treated as the condition under which an MNC subsidiary unit gains an edge over a typical rival firm of the host country. Now λ, the challenge cost for the MNC per production centre is plausibly much insignificant in relation to β, the value of the knowledge-based assets. Moreover, justifiably enough, $\lambda\to\epsilon$ in the long run, where ϵ is indeed a very small quantity (>0). This is so because, as argued earlier, with passage of time the production centre of the MNC in every host country gets along with the challenges with lesser amount of difficulties. Hence the result of a substantial cost advantage for an MNC unit over an RRU.

Conclusion

The preceding analysis is a strong pointer to the reason behind the dominance of the MNCs over the local firms in the host countries. It is simply the public good property of the knowledge-based assets that enables an MNC to attain a substantial cost advantage over the local firms. The rival firms are deprived of the cost advantage only because their scale of operations is much on the lower side vis-a-vis an MNC, as clearly shown above. Our analysis further indicates that larger is the global spread or representation of an MNC, higher is the degree of its cost advantage over the host country firms.

References

Dunning, J (1977). 'Trade, Location of Economic Activity and MNE : A Search for an Eclectic Approach'.

---------- (1981). International Production and the Multinational Enterprise. London : George Allen and Unwin.

Ethier, W.J. (1986). "The Multinational Firm". Quarterly Jounrnal of Economics 80:805-833.

A Primer on Handling Unusual Observations in Quantitative Data Analysis

Sharmila Banerjee,
Reader, Department of Commerce,
University of Calcutta

Abstract

Data analysis in any business setting, involves the application of appropriate techniques for extracting the information contained in the data. Statistical methods are considered as valuable tools in this context. Unusual observations in any dataset can often mislead the analysis and influence the outcome of the study. These effects of outliers have been illustrated with simple examples. Two basic approaches and corresponding methods of handling outliers have been discussed.

Key-Words: Outlier; Boxplot; Robust; Accommodation; Identification

1. Introduction

The presence of some anomalous or doubtful observations in a dataset has always demanded some special attention. In many observational studies involving data collection, it is an unfortunate fact that data are not always well behaved. Data may have unusual values or outliers, even if they come from reliable sources. Like other research studies, in market research, business forecasting and decision-making processes, these unusual observations should be handled with utmost care in order to have meaningful results. Researchers and data analysts, working with genetic, environmental or commercial data, are encountering large data sets in which the problem becomes more acute. For these reasons, it is essential to understand various aspects of outliers, e.g. the meaning and nature of outliers, their sources, and means of handling them. So it is very useful to have simple and effective graphical summaries, summary statistics, and screening procedures that can easily identify and handle these outliers.

2. What is an Outlier and Why Do Outlying Observations Arise

In the past, outliers were only viewed as observations "which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which in to the investigation can only serve... to perplex and mislead the enquirer." (Source: Barnett and Lewis,1994). In the course of time, this concept of defining outliers has changed. Outliers in a dataset do not always appear as errors or contaminants, rather sometimes they might represent some new sources of observations, which are also to be considered in the investigation. Therefore, one objective of handling the outliers is to see that the conclusions drawn from the analysis are not influenced by these doubtful or extreme observations. At the same time a new thought on the data source or data-generating mechanism is also explored.

The variability in a dataset can be attributed to several sources. When a sample data is taken, it is natural that the observations will vary over the population. The inherent variability

in the sample observations reflects the distributional properties of the population. Inadequacies in measuring instruments additionally contribute to the inherent variability. Error in selection of samples also gives rise to another source of variability that is not contemplated in the analysis. Biased sample selection may violate the basic assumption about the population distribution. As all outliers are not contaminants, outliers in a dataset may in fact be a perfectly reasonable reflection of the natural inherent variation. Though they seem statistically unreasonable, in reality the assumed basic data generating model may be inadequate.

3. Problematic Effects of Outliers

A very important part of a thorough analysis is to look for the unusual observations or outliers and to understand how they impact data analysis. Presence of outliers in any dataset can affect the outcome of the analysis in several ways. It generates bias in the estimation of population parameters. In some extreme cases the bias is so high that the estimates get distorted. Table 1 illustrates a very simple example, which gives an idea about the potential problems that arise due to irregular or unusual observations.

Table 1

	Sorted data	Mean	Median	Variance	95% confidence interval for the mean
Data without irregular/unusual observation	2,3,5,7,9,11, 11.12.14	8.22	9	17.69	(4.99,11.46)
Data with one irregular/unusual observation (marked by *)	2,3,5,7,9,11, 11,12,140*	22.22	9	1963.7	(-11.8,56.3)

Even in simple testing of hypothesis for the population mean, the presence of outliers can lead to wrong conclusion. It will change the type I error and the power of the test significantly. Table 2 gives an idea about how a single outlier can reduce the power of test for the mean. In testing of mean for samples of size 50, 100, 1000 from a normal population with standard deviation 1 and nonzero mean (μ_1), the minimum values of μ_1 which can be detected as nonzero, with 80% power are .3572, .2506, and .0787, respectively. From Table 2 we see that even one outlier (having value 10) can reduce the power of the test substantially.

Table 2 : Power of test for mean.

Sample size	Power (No outlier)	Power (1 outlier having value 10)
50	80%	2.8 %
100	80%	18.1%
1000	80%	67.3%

In regression analysis the estimated regression coefficients that minimize the Sum of Squares for Error (SSE) are very sensitive to outliers. In bivariate regression analysis, while studying the relationship between two variables, the presence of unusual observations can greatly influence the outcome of the study. They distort the estimates of intercept and slope parameters and certainly inflate their standard errors. Special attention is required for the observations whose x values (the predictor variable values) are far from the mean value because they have greater influence on the regression than the influence of those, which are nearer. In analysis of variance, presence of unusual observations results in inflated sum of squares, and as a consequence partitioning the sources of variation in the data in to some meaningful components becomes difficult. Figure 1 and Table 4 show the lines of fit and the regression analysis results for the following data on the number of employees, x, and the profits per employee, y, for n=16 publishing firms (Forbes, April 30, 1990) in Table 3.

Table 3

x (1000's employees)	9.4	6.3	10.7	7.4	17.1	21.2	36.8	28.5
y (1000's dollars)	33.5	31.4	25.0	23.1	14.2	11.7	10.8	10.5

x (1000's employees)	10.7	9.9	26.1	70.5*	14.8	21.3	14.6	26.8
y (1000's dollars)	9.8	9.1	8.5	18.3*	4.8	3.2	2.7	-9.5

The first plot in Figure 1 includes all the observations and the 2nd plot is obtained by eliminating one observation (marked by * in Table 3), which seems separated compared to the rest of the observations. It is very clear from the plots and the results in Table 4 that the regression coefficients, the standard error (StDev), and therefore the regression line are greatly influenced by this observation. Inclusion of this observation in the study makes the model weak and less effective.

Table 4

Regression Analysis (Considering all observations)

The regression equation is
y = 18.0 - 0.271 x

Predictor	Coef	StDev	T	P
Constant	17.954	4.457	4.03	0.001
x	–0.2715	0.1726	–1.57	0.138

S = 10.61 R–Sq = 15.0% R–Sq(adj) = 9.0%

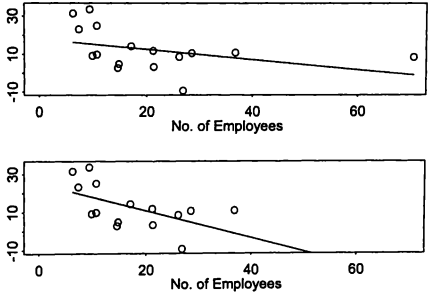
Regression Analysis (Eliminating the unusual observation)

The regression equation is
y = 25.0 - 0.713 x

Predictor	Coef	StDev	T	P
Constant	25.013	5.679	4.40	0.001
x	–0.7125	0.2912	–2.45	0.029

S = 9.839 R–Sq = 31.5% R–Sq(adj) = 26.3%

Figure 1



While studying predictability of stock market returns using time series analysis, the results can be biased because of infrequent outliers in the data. They are detected as observations having large residuals in an estimated autoregressive model. The results improve substantially if these outliers are taken into account in making one- step ahead predictions.

4. Some Basic Methods of Handling Outliers

The primary two objectives of handling outliers are to identify the potential outliers, and to accommodate them so that their influence on the analysis is minimized. The accommodation approach is based on the use of robust methods that reduce the influence of outliers. Instead of discarding the observations, which seem to be outliers, these methods use them in a way so that their influence on the inference and analysis is minimized. Here the main aim is to accommodate outliers, that is, "to play safe against their potential dangers and to render their effects in the overall result harmless," Hampel (1968). In recent years major efforts have been made to obtain statistical procedures, which provide a measure of protection against various types of uncertainty of knowledge of the data generating mechanism. Simple, informative and detailed texts on robust statistics are provided by Huber (1981), Rousseeuw and Leroy (1987), Tukey (1960), and Hampel (1974). Median is a robust estimator of the central tendency in the sense that it is less influenced by the extreme values. Trimmed mean and Winsorized mean are two commonly used robust measures of central tendency. While obtaining the estimates they put less weight for the extreme values (the higher and lower end values in ordered data) than the other observations. Trimmed mean of a data set is the simple arithmetic mean of the trimmed data, where the k1 lowest and k2 highest (for suitably chosen k1 and k2) extreme values are ignored. In Winsorization, k1 lowest and k2 highest extreme values are replaced, respectively, by the lowest and highest values of the remaining observations. The mean of the new set of observations is called the Winsorized mean.

The Annual Respondents Database (ARD) which stores the data collected by the office for National Statistics (formerly the Central Statistical Office) from the Annual Census of Production and Annual Census of Construction in UK, regards a response as "an outlier if it is outside certain limit of what to expect for that enterprise, i.e. when compared with previous surveys or administrative data. Where this cannot be reconciled through follow up enquiries, smoothed outliers are added to the original and constructed respondent data to produce a set of variable sets prefaced by 'WQ'. The 'W' refers to the method of smoothing outliers known as Winsorization" (ref : www.statistics.gov.uk).

The basic problem in trimmed and Winsorized mean is choosing the extent of trimming or Winsorization. Methods have been developed for trimming or Winsorizing in terms of some quantitative measures of their extremeness, rather than just considering their position in the ordered data. Some examples are modified trimming (Anscombe,1960), modified Winsorisation (Guttman and Smith, 1969) and semi Winsorisation (Guttman and Smith , 1969).

Standard deviation, as an estimator for spread is very sensitive to outliers, and therefore not robust. Inter quartile range (IQR), Median absolute deviation (MAD) are simple robust measures of spread or variation in the data. IQR is the difference between Q_3 and Q_1, where Q_3 and Q_1 are the 3rd and 1st quartile, and MAD = median $|$ xi - median(x)$|$.

Outlier identification methods basically label each observation in a sample as outside or inside some specific interval and accordingly identify it as an outlier or a regular observation. For identification of outliers, one exploratory data analytic tool called boxplot has found great popularity. It is a graphical univariate data summary. It consists of the median, the lower and upper quartiles and the smallest and largest observations. In Figure 2 a typical boxplot has been constructed for the data in Table 5. It is represented by a box where the horizontal central line within the box represents the median = Q_2 = 11.947, the upper and lower horizontal lines enclosing the box are at the upper and lower quartiles, Q_3 = 21.571, and Q_1 = 3.692, respectively. The boxplot shows the data center through Q_2 , the variability by IQR which is equal to $Q_3 - Q_1$, and the skewness (asymmetry) through the differences $Q_3 - Q_2$, and $Q_2 - Q_1$. Like Q_2 (median) , Q_1 and Q_3 are also robust estimators, and are not unduly affected by a few unusual observations. 50% of the observations lie between Q_1 and Q_3. The boxplot modified by Tukey (1977) uses the simple outlier labeling rule that flags observations as outliers if they fall above the upper fence or below the lower fence, where the upper fence equals $Q_3 + k(Q_3 - Q_1)$, and the lower fence equals $Q_1 - k(Q_3 - Q_1)$. The suggested value of k is 1.5. In boxplot the whiskers extend up to the most extreme values within the fences. Any value outside the fence will be identified as outlier. The minimum and the maximum of the observations in the data are, .707 and 51.284 respectively, and 1.5 X IQR = 26.819. The minimum observation in the data set, .707 is well within 26.819 of Q_1= 3.692. Therefore the left hand whisker extends up to this smallest value. Q_3 + 1.5 X IQR = 48.390; so the right hand whisker extends to the largest observation in the dataset less than or equal to 48.390 (here 38.173). The only observation 52.284, which is outside the upper fence has been plotted individually, and it is highlighted (marked as *) as a project whose construction cost is significantly higher than the others. Though most of the statistical data analysis software use k =1.5 to flag the outliers, for large data sets it becomes ineffective and the value of k needs revision. Hoaglin and Iglewicz (1987), and Banerjee and Iglewicz (2001) suggested values for k for different sample size. For skewed

data, Kimber (1990) suggested a simple adjustment in which the lengths of whiskers are adjusted for asymmetry. In some cases skewed data can be made symmetric by a simple transformation (e.g. Log transformation). In that case, ordinary boxplot will serve the purpose.

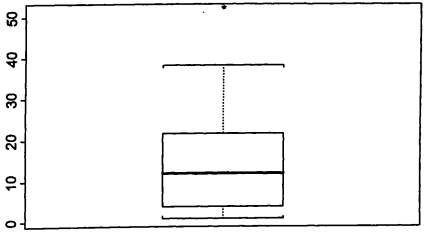
Table 5

Data on actual costs in millions of dollars of 26 construction projects at a large industrial facility (source: Schmoyer, 1992).

.918	7.214	14.577	30.028	38.173	15.320
14.837	51.284	34.100	2.003	20.099	4.324
10.523	13.371	1.553	4.069	27.973	7.642
3.692	29.522	15.317	5.292	.707	1.246
1.143	21.571				

Figure 2

Boxplot for Construction Cost

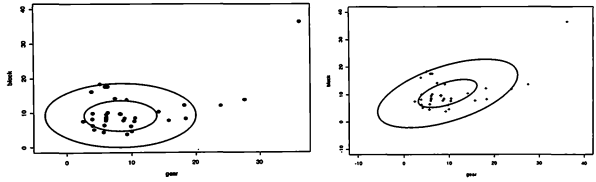


There are statistical tests to check whether an extreme observation is an outlier. These tests require the assumptions on the data generating model, and are aimed at testing an extreme observation with the prospect of rejecting it from the data set or identifying it as a feature of special interest. A statistical test is performed on the extreme observations to examine whether they are not only extreme but also statistically unreasonable, even when viewed as extreme. The test statistic is usually of the form N/D where the numerator N is a measure of the separation of the extreme observation(s) from the remainder of the sample and the denominator D is a measure of the spread of the sample. A number of these tests and corresponding tables are available in Barnett and Lewis (1994).

5. Outliers in Bivariate data

There are number of plots available to visualize bivariate data. Scatter plots, local density plots, smoothed curve scatter plots are some of them. The idea of boxplot for univariate data has been extended to bivariate data, and is capable of identifying unusual observations. Bivariate normal probability plot is the basis of a bivariate boxplot. Bivariate normal probability plot requires five estimated parameters; the two means μ_x, μ_y; the two standard deviations σ_x, σ_y; and the correlation coefficient ρ. The inner ellipse consists 50% of the probability plot and outer ellipse consists 90% of the probability plot. As the estimators are not robust, the direction and the shape of the distribution indicated by the plot are influenced by the extreme observations and the plot fails to detect the true outliers. Relplot (Robust Elliptical Plot) and Quelplot (Quarter Elliptic Plot) are two robust plots devised by Goldberg and Iglewicz, in which the robust estimates are used to obtain the plots. Figure 3 shows bivariate normal probability plot (1st plot in Figure 3), and robust elliptical plot (2nd plot in Figure 3), for the same data. Relplot identifies two outliers.

Figure 3

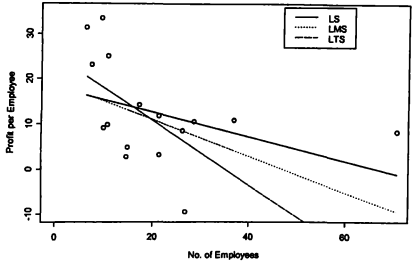


In the relplot, the outer ellipse is drawn in such a way, that the outlier labeling rule for the plot has similar outlier detection probabilities as the univariate boxplot with standard 1.5 multiple of the interquartile range. The relplot contains a smaller area than the bivariate normal plot but its inner ellipse contains more points than the bivariate normal plot. The direction also differs in these two plots. This difference is due to the influence of the farthest northeast points on the normal probability plot.

In bivariate regression analysis the influence of extreme observations has already been explained in Table 3 and Figure 1. These outlying observations are often hidden by the fitting process and may not be easily detected. Yet they can have a major influence in determining the fitted regression function. Many software packages consider standardized residuals or studentized residuals to detect unusual observations. The weighted or studentized residuals provide an appealing measure of the deviation of each observation from the trend expressed by the other observations in the data. Observations with large standardized residuals (statistically significant) are typically identified by the statistical packages as unusual observations. The observation marked by * in Table 3 has large studentized residual and can be detected (at 5% level) as unusual observation by standard statistical packages.

As an attempt to get a regression fit which is less influenced by extreme observations, Rousseeuw (1984) suggested least median of square (LMS) regression. Rather than minimizing the sum of squared residuals, he proposes minimizing the median of the squared residuals. Another alternative suggested by him is least trimmed square (LTS) regression, where the sum of smallest q residuals (for suitably chosen q) is minimized. It was found that the estimators found in both cases were remarkably resistant to a number of contaminants. Figure 4 shows the LTS and the LMS regression line along with the ordinary least square (LS) line for the data considered in the analysis in Table 3.

Figure 4



Statistical software packages like SAS, MINITAB, S-plus etc. have options to handle outliers. All these software packages have codes and visual tools to generate Boxplots. For regression analysis, they all offer diagnostics to mark observations that have high levels of studentized residuals. SAS and S-plus also provide robust regression methods. Codes for constructing relplots are available in S-Plus.

6. Summary and Conclusion

The correct measure of extremeness in any observation is always relative to proper assumption of the underlying data-generating model. An unusual observation may be a contaminant or may be a regular observation, which deserves special attention. It is important to study outlying observations to decide whether they should be retained or eliminated, and if retained whether their influence on the analysis should be reduced. There are robust estimators and robust statistical procedures, which accommodate the outliers and perform well in a variety of settings. In many of these procedures the assumptions on the data-generating model can be relaxed. Trimmed mean, Winsorized mean, IQR, LMS and LTS regressions are some of the examples of robust methods of estimation and analysis. Outlier identification rules have been designed

and extended to large samples. Boxplots, Relplots etc. are outlier identification tools for univariate and bivariate samples. Standard statistical software packages have features to address problems related to handling of outliers.

References

Anscombe, F. J. (1960) "Rejection of Outliers", Technometrics 2, 123-147.

Banerjee, S. and Iglewicz, B. (2001) "A simple Univariate Outlier Identification Procedure," American Statistical Association Proceedings., August 2001.

Barnett, V., and Lewis, T. (1994) *Outliers in Statistical Data (3rd ed.)*, New York : Wiley.

Goldberg, K. M., and Iglewicz, B. (1992) "Bivariate Extensions of the Boxplot," Technometrics., 34, 307-320.

Guttman, I., and Smith, D. E. (1969) " Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution I: Estimation of Mean," Technometrics. 11, 527-550.

Hampel, F. R. (1974) "The Influence Curve and its Role on Robust Estimation," Journal of American Statistical Association.,69, 389-393.

Hampel, F. R. (1968) "Contributions to the theory of Robust Estimation," Ph. D. thesis, University of California, Berkley.

Hoaglin, D. C., and Iglewicz, B. (1987) "Fine Tuning Some Resistant Rules for Outlier Labeling," Journal of American Statistical Association., 82, 1147-1149.

Huber, P. J. (1981) *Robust Statistics.*, New York: John Wiley.

Kimber, A. C. (1990) "Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions," Applied Statistics., 39, 21-30.

Rousseeuw, P. J., and Yohai, V. (1984) "Robust Regression by means of S-estimators. Lec. Notes Statist.," 26, 256-272.

Rousseeuw, P. J., and Leroy, A. M. (1987) *Robust Regression and Outlier Detection.*, New York : John Wiley.

Scmoyer, R. L.(1992) "Asymptotically Valid Prediction Intervals for Linear Models," Technometrics, 34, 399-408.

Tukey, J. W. (1960) *A survey of Sampling from Contaminated Distributions*, University Press, Stanford, California.

Tukey, J. W. (1977) Exploratory Data Analysis, Addison-Wesley, Reading.