STATISTICAL ANALYSIS/METHODS OF DETECTING OUTLIERS IN A BIVARIATE DATA IN A REGRESSION ANALYSIS MODEL

EKEZIE DAN DAN AND OGU AGATHA IJEOMA

Department of Statistics
Imo State University, PMB 2000, Owerri Nigeria

ABSTRACT

This study detects outliers in a bivariate data by using identification of influential observation and scatter diagram. The study shows how an observation that causes the least squares point estimates of a Regression model to be substantially different from what it would be if the observation were removed from the data set. A Boilers data with a dependent variable Y (Man-Hours) and four independent variables X_1 (Boiler Capacity), X_2 (Design Pressure), X_3 (Boiler Type), X_4 (Drum Type) were used. The analysis of the Boilers data reviewed an unexpected group of outliers. MINITAB (version 11.0) software was used to analyze all the regression models. Microsoft Excel (version 2003) software was used in plotting the scatter plots. The results from the findings showed that an observation can be outlying with respect to its Y (dependent) value or X (independent) value or both values and yet influential to the data set.

Keywords: Detecting Outliers, Scatter Diagram, Identification of Influential observation and Regression Analysis.

1 INTRODUCTION

"Outliers" are unusual data values that occur almost in all research projects involving data collection. This is especially true in observational studies where data naturally take on very unusual values, even if they come from reliable sources. Although definitions varies. Outliers are observations that have extreme value relations. An outlier is generally considered to be a data point that is far outside the norm for a variable or population Jarrell, (1994); Rasmussen, (1988) and Stevens (1984).

Hawkin described an outlier as an observation that "deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Outliers have also been defined as values that are "dubious in the eyes of the researcher" Dixon, (1950) and contaminants Wainer, (1976).

In the presence of outliers, any statistical test based on sample means and variance can be distorted. For instance, estimated regression coefficients that minimize the Sum of Squares for Error (SSE) are very sensitive to outliers. There are several problematic effects of outliers which include:

(a) Bias or distortion of estimates
(b) Inflated sum of squares (which make it unlikely to partition the source of variation in the data into meaningful components).
(c) Distortion of p-values (statistical significance, or lack thereof can be due to the presence of a few-or even one-unusual data value).
(d) Faulty conclusions (it's quite possible to draw false conclusions if you haven't looked for indications that, there was any thing unusual in the data). Thus, the need of screening data for Univariate, bivariate and multivariate outliers is important in these days of ubiquitous computing.

2 CAUSES OF OUTLIERS

Outliers can arise from several different mechanisms or causes. Anscombe (1960) sorts outliers into two major categories: those arising from errors in the data and those arising from the inherent variability of the data.

NOTE: Not all outliers are illegitimate contaminants and not all illegitimate scores show up as outliers, Barnett and Lewis, (1994). It is therefore important to consider the range of causes that may be responsible for outliers in a given set of data.

- Outliers from data errors: outliers are often caused by human error, such as errors in data collection, recording or entry. Data from an interview can be recorded incorrectly, or mistaken upon data entry. Errors of this nature can often be corrected by returning to the original documents or even the subjects if necessary and possible and entering the correct value.
- Outliers from intentional or motivated mis-reporting: There are times when participants purposefully report incorrect data to experimenters or surveyors. A participant may make conscious effort to sabotage the research Huck, (2000), or may be acting from other motives. Social desirability and self-presentation motives can be powerful. This can also happen for obvious reasons when data are sensitive (e.g. teenagers under-reporting drug or alcohol use, misreporting of sexual behaviour). If all but few teens under-report a behaviour (for example, the frequency of sexual fantasies teenage male experience) the few honest responses might appear to be outliers when in fact they are legitimate and valid scores. Motivated over-reporting can occur when the variable in question is socially desirable (e.g. income,

educational attainment, grades, study times, church attendance, sexual experience). Environmental conditions can motivate over-reporting or mis-reporting, such as if an attractive female researcher is interviewing male undergraduates about attitude on gender equality in marriage. Depending on the details of the research, one of two things can happen: inflation of all estimates, or production of outliers. If all subjects respond the same way, the distribution will shift upward, not generally causing outliers. However, if only a small subsample of the group responds this way to the experimenter, or if multiple researchers conduct interviews, then outliers can be created.

- Outliers from sampling error: another cause of outliers is sampling. It is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample. For example, in education, in advert entry sampling academically gifted or mentally retorted students is a possibility and (depending on the goal of the study) might provide undesirable outliers. These cases should be removed as they do not reflect the target population.
- Outliers from standardization failure: outliers can be caused by research methodology, particularly if something anomalous happened during a particular subject experience one might argue that a study of stress levels in school children around the country might have found some significant outliers. Unusual phenomena such as a construction noise outside a research laboratory or an experimenter feeling particularly grouchy, or even events outside the context of the research laboratory, such as a student protest, a rape or murder on campus, observations in the classroom the day before a big holiday recess and so on can produce outliers. Faulty or non-calibrated equipments is another common cause of outliers. These data can be legitimately discarded if the researchers are not interested in studying the particular phenomenon in question (e.g. if I were not interested in studying my subjects' reactions to construction noise outside the laboratory).
- Outliers from faulty distributed assumptions: incorrect assumptions about the distribution of the data can also lead to the presence of suspected outliers Iglewieze and Hoaglin, (1993). Blood sugar levels, disciplinary referrals, scores on classroom tests where students are well-prepared, and self-reports of low-frequency behaviours (e.g. number of times a student has been suspended or held back a grade) may give rise to bimodal, skewed, asymptotic or flat distributions, depending upon the sampling design. The data may have a different structure than the researcher originally assumed, and long or short-term trends may affect the data in unanticipated ways. For example, a study of college library usage rates during the month of September may find outlying values at the beginning and end of the month, with exceptionally low rates at the beginning of the month when students have just returned to campus or are on break for labour weekend in (Nigeria) and exceptionally high rates at the end of the month, when mid-term examinations have begun. Depending on the goal of the research, these extreme values may or may not represent an aspect of the inherent

variability of the data, and may have a legitimate place in the data set.

- Outliers as legitimate cases sampled from the correct population: it is possible that an outlier can come from the population being sampled legitimately through random chance, it is important to note that sample size plays a role in the probability of outlying values. Within a normally distributed population, it is more probable that a given data point will be drawn from the most densely concentrated area of the distribution, rather than one of the tails Evans, (1999); Sachs, (1982). As a researcher casts a wider net and the data set becomes larger, the more the sample resembles the population from which it was drawn and thus the likelihood of outlying values become greater. In other words, there is only about one percentage chance you will get an outlying data point from a normally distributed population, this means that, on the average, about one percentage of your subjects should be three standard deviations from the mean. In the case that outliers occur as a function of the inherent variability of the data, opinions differ widely on what to do. Due to the dexterous effect on power, accuracy and error rates that outliers can have, here it might be desirable to use a transformation or recoding/truncation strategy to both keep the individual in the data set and at the same time minimize the harm to statistical inference: Osborne, (2002).
- Outliers as potential focus of inquiry: we all know that interesting research is often as much a matter of serendipity as planning and inspiration. Outliers can represent a nuisance error, or legitimate data. They can also be inspiration for inquiry. When researchers in Africa discovered that some women were living with HIV just fine for years and years, untreated, those rare cases were outliers compared to most untreated women, who die fairly rapidly. They could have been discarded as noise or error, but instead they serve as inspiration for inquiry. This extreme score might shed light on an important principal or issue. Before discarding outliers, researchers need to consider whether those data contain valuable information that may not necessarily relate to the intended study, but has importance in a more global sense.

The presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistics estimates when using either parametric or nonparametric tests (Zimmerman, 1994, 1995, 1998). Casual observation of the literature suggests that researchers rarely report checking for outliers of any sort. This inference is supported empirically by Osborne, Christiansen and Gunter (2001), who found that authors reported testing assumptions of the statistical procedure(s) used in their studies – including checking for the presence of outliers – only eight per cent of the time. Given what we know of the importance of assumptions to accuracy of estimates and error rates, this in itself is alarming.

Wainer (1976) also introduced the concept of the “froigelier” referring to “unusual events which occur more often than Seldom” (p. 286). These points lie near three standard deviations from the mean and hence may have a disproportionately strong influence on parameter estimates, yet are not so obvious or easily identified as ordinary outliers due to their relative proximity to the distribution center. As fringeliers are a special case of outliers, for much of the rest of this study we will use the generic term “outlier” to refer to any single data point of dubious origin or disproportionate influence.

Hadi and Simonoff (1993) provided distributional results for testing, multiple outliers in regression analysis. The test is based on the deletion residual. Beckman and Cook (1983) encountered a serious problem of "masking" if there are several outliers. Least square estimation of the parameter of the model may lead to small residuals for the outlying observations. Single detection methods (for example Cook and Weisberg, 1982; Alkinson, 1985) may fail and the outliers will go undetected.

Hawkins (1983) argues for exclusion of all possible outlying observations, which are then tested sequentially for reinclusion. The drawback to this procedure is that it is unclear how many observations should be deleted, and because of masking, which ones, before reinclusion and testing begin.

The use of the forward search in regression is described in Atkinson and Riani (2000) whereas in Atkinson (1994) the emphasis on informative plots and their interpretations. Although the forward search is a powerful general method for the detection of multiple outliers and unidentified clusters and of their influential effects. The interest here is in Atkinson (1994) on information plots and the information it provides about the adequacy of our simple approximation to the distribution of the test statistic.

Possible sources of outliers are recording and measurement errors is correct distribution assumption unknown data structure, or novel phenomenon (Iglewiez, 1993). A data set indicative of a novel phenomenon can be often labeled as an outlier. For instance, the measurements indicating existence of the hole in the ozone layer were initially thought to be outliers and they were automatically discarded. This join delayed the discovery of the phenomenon by several years (Berthouex, 1994). The first step in data analysis is to label suspected outliers for further study. Three different methods are available to the investigation for normally distributed data: z-score method, (Iglewiez, 1993; Barnett, 1984). All of the experimental observations are standardized and the standardized values outside a predetermined bound are labeled as outliers (Rousseeuw, 1987).

Outliers can arise from several different mechanisms as causes. Anscombe (1960) sorts outliers into categories from intentional or motivated misreporting; a participant may make a conscious effort to sabotage the research (Huck, 2000) or may be acting from other motives. In outliers from faulty distributional assumptions, incorrect assumption about the distribution of the data can also lead to the presence of suspected outliers (Iglewiez and Hoaglin, 1993). Due to the deleterious effect on power accuracy, and error rates that outliers can have, it might be desirable to use a transformation or recording strategy to both keep the individual in the data set and at the same time minimize the harm to statistical inference (Osborne, 2002).

Rosner's Test identifies outliers that are both high and low; it is therefore always two tailed (Gibbon, 1994). The R. Statistics is compared with a critical value (Gilbert, 1987). Rosner's (1983) "many outlier" sequential procedures is an improved version of Rosner's (1983) "extreme studentized deviate" outlier test. Simonoff (1982) found this earlier well compared to other outlier test, although Rosner (1983) points out that it tends to detect more outliers than are actually present. Rosner's (1983) method assumes that the main body of data is from a normal distribution.

Rosner's tests are two tailed since the procedure identifies either suspiciously large or suspiciously small data. When a one tailed test is needed, that is when there is interest in detecting only large values or only small values, then the skewness test for outliers

discussed by Barnett and Lewis (1994) is suitable.

Hamilton, L.C. (1982) give a graphical procedure for identifying outliers from bivariate normal or bivariate log normal distributions.

Rather than identifying outliers and discarding them before doing least square regression, one could do robust regression, as discussed and illustrated by Rousseeuw and Leroy (1987) caution that robust regression should be applied only after the investigator is satisfied that less weight should be applied to the divergent data. Non-parametric regression discussed by Holander and Wolfe (1973), and Reckhow and Chapra (1983) is an alternative to either standard least squares regression or robust regression.

Methods for detecting outliers have received a great deal of attention recently Cook and Wainer, 1976 and Steven, 1984). Leverages are related to an alternate regression diagnostic, Mahalanobis distance (Stevens, 1984).
Mixture regression occurs when there is an omitted categorical predictor like gender, species or location and different regression occur in each category. It has long been recognised that a lurking variable, a variable that has an important effect but is not present among the predictors under consideration (Box, 1966; Joiner, 1981; Moore, 1997) can complicate regression analyses.

Atkinson, (1994) have applied Akaike Criterion (AIC) in detection of outliers by using (quasi) Bayesian approach with predictive likelihood in place of the usual likelihood function otherwise, detection of outliers has a long history. The main theme, however, has been around univariate and single outliers. Recently, some promising results have been obtained in detecting multiple outliers also in multivariate analysis (Hadi, 1992).

An approach to the identification of aberrant points is the construction of outliers' diagnostics. These are quantities computed from the data with the purpose of pinpointing influential points, after which these outliers are to be removed or corrected, followed by a least square analysis on the remaining cases. When there is only a single outlier, some of these methods work quite well by looking at the effect of deleting one point at a time. (Atkinson, 1985;) Cook and Weisberg, 1982 and Hawkins, 1980). Unfortunately, it is much more difficult to diagnose outliers when there are several of them, due to the so-called masking effect which says that one may mask another. The naira extensions of classical diagnostics to such multiple outliers often give rise to extensive computations. Recent work by Atkinson (1986), Hawkins, Bradu and Kass (1984), and Rousseeuw and Van Zomeran (1999) indicates that one needs to use robust methods in one way or another to safely identify multiple outliers.

Some researchers prefer visual inspection of the data. Others (Lornez, 1987) argue that outlier detection is merely a special case of the examination of data for influential data points. In analysis of variance, the biggest issue after screening for univariate outliers is the issue within cell outliers or the distance of an individual from the subgroup. Standardised residuals represent the distance from the subgroup and thus are effective in assisting analysis in examining data for multivariate outliers. Tabachnick and Fidell (2000) discuss data cleaning in the context of other analyses.

Where outliers are illegitimately included in the data, it is only common sense that those data points should be removed (Barnett and Lewis, 1994). Few should disagree with that statement. When the outlier is either a legitimate part of the data or the cause is

unclear, the issue becomes unclear. Murkier Judd and McClelland (1989) make several strong points for removal even in these cases in order to get the most honest estimate of population parameters possible (Barnett and Lewis, 1994).

On means of accommodating outliers is the use of transformations (Osborne, 2002). By using transformation extreme scores can be kept in the data set, and the relative ranking of scores remains yet the skew and error variance present in the variable can be recorded (Hamilton, 1992).

However, transformations may not be appropriate for the model being tested or may affect its interpretation in undesirable ways. Taking the logarithms of a variable makes a distribution less skewed, but it also alters the relationship between the original variables in the model (Newton and Rudestam, 1999; Osborne, 2001).

Instead of transformation, researchers sometimes use various robust procedures to protect their data from being distorted by the presence of outliers. These techniques "accommodate the outliers at no serious inconvenience or are robust against the presence of outliers (Barnett and Lewis, 1994; p. 35). Certain parameter estimates, especially the mean and least square estimates, are particularly vulnerable to outliers, or have "low breakdown" values. For this reason, researchers turn to robust or "high breakdown" methods to provide alternative estimates for these important aspects of data.

A common robust estimation method of the univariate distributions involves the use of trimmed mean, which is calculated by temporarily eliminating extreme observations of both ends of the sample (Anscombe, 1960). Alternatively, researchers may choose to compute a winsorized mean, for which the highest and lowest observations are temporarily censored, and replaced with adjacent values from the remaining data (Barnett and Lewis, 1994).

Assuming that the distribution of prediction errors is close to normal, several common robust regression techniques can help reduce the influence of outlying data points. The least trimmed squares (LTS) and the least median of squares (LMS) estimators are conceptually similar to the trimmed mean, helping to minimize the scatter of the prediction errors by eliminating a specific percentage of the largest positive and negative outliers (Rousseeuw and Leroy, 1987). While Winsorized regression smoothes Y-data by replacing extreme residuals with the next closest value in the dataset (Lane, 2002).

In correlations, we are expected to see the effect of outliers on two different types of correlations. These are correlations close to zero (to demonstrate the effect of outliers on Type II error rates) correlations will be calculated in each sample both before removal of outliers and after. If a sample correlation leads to a decision that deviated from the "correct" state of affairs it was considered an error or inference. In most cases the incidence of errors of inference was lower with cleaned than unclean data.

For the T-test and Analysis of Variance (ANOVA) this deals with analysis that look at group mean differences, such as the t-test and analysis of variance. For the purpose of simplicity these analyses are simple t-tests but these results would be generalized to any analysis of variance. For these analyses two different conditions are examined when there were no significant differences between the groups in the population and when there were significant group differences in the population. For both variables the effects of having outliers in only one cell as compared to both cells were examined.

Removal of outliers will produce a significant change in the mean differences between two groups. It will also produce significant change in the t-statistics. Evidence of outliers may produce type I or type II errors. Removal of outliers may tend to have a significant beneficial effect on error rates.

Most analysts argue that removal of extreme scores produces undesirable outcomes; they are in the minority especially when the outliers are illegitimate. When the data points are suspected of being legitimate, some authors Orr, Sacketts, P.R. and DuBois (1991), argue that data are more likely to be representative of the population as a whole if outliers are not removed.

Conceptually, there are strong arguments for removal or alteration of outliers. In some analyses the benefits of outliers' removal are reported. Both correlations and t-tests may show significant changes in statistics as a function of removal of outliers. In most cases errors of inference were significantly reduced, a prime argument for screening and removal of outliers. It is straightforward to argue that the benefits of data cleaning extend to simple and multiple regressions to different types of ANOVA procedures. There are other procedures outside these but the majority of social science research utilizes one of these procedures. Other researches (e.g. Zimmerman, 1995) have dealt with the effects of extreme scores in less commonly used procedures, such as nonparametric analyses. Thus, checking for the presence of outliers and understanding how they impact data analysis are extremely part of a complete analysis, especially when any statistical technique is involved.

This study will examine the causes, problems, methods of detection and approaches to data analysis of outlier in a Univariate, Bivariate and Multivariate data using four test methods namely; Rosners', Grubbs', Data plots and Leverage approach in a regression analysis model.

3 IDENTIFICATION OF OUTLIERS.

There is no such thing as a simple test. However, there are many ways to look at a distribution of numerical values, to see if certain points seem out of line with the majority of the data. And expert knowledge of what values data can have is probably the best solution. Thus, there are some guidelines with which one can always begin.

The "normal" distribution myth. Although not necessarily an issue with outliers, it is important to first recognize what the distribution of your data looks like. For many statistical modeling purposes, the data do not require a "normal" or symmetric, bell-shaped distribution. (This assumption applies to the residuals from a liner statistical model). Data collected as counts will not usually look very "normal". Data that are collected across group may have a distribution that has several local peaks. In fact, for data to be entered into a linear regression model, it is preferable for the independent or explanatory variables to not have a normal distribution. The mathematics behind linear regression demonstrates that normality is not required or even desirable for this type of analysis. What is important is to check for data values that lie well out side the range of other data called "leverage points" that will likely exert a strong influence on the results. The objective is to collect data with a distribution that allows one make the best influence possible about the population under study.

- Visual Aids: Always check the distributions of data whether they be nominal or continuous. This procedure should be one of the first steps in data analysis as it will quickly reveal the most obvious outliers. For continuous or interval data, a dot plot of a single variable or multi-dimensional of all

pair wise scatter plots of continuous variables are good methods to visually detect outlying observations. With larger sample sizes a box plot is another very helpful tool, since it makes no distributional assumptions which are often not relevant (e.g. assume a normal distribution when you may have skewed non-negative data). They also may require that a location (mean) or scale (standard deviation) parameter be estimated from the data. As said earlier, outliers greatly influence these two summary statistics. This is one reason why eliminating data that exceed two or three standard deviations may not be a good or even a reasonable decision rule.

- IQR Computation: a simple task is to compute the inter-quarter-range (IQR) for continuous data and then take a multiple of it as a cut-off value to define values which are considered outliers. For large datasets, a box plot applies this technique to identify outliers. It is an extremely effective approach, especially when you have thirty or more data points within each group level.

4 DEALING WITH OUTLIERS.

There is a great deal of debates as to what to do with identified outliers. A thorough review of the various arguments is not possible here rather will be seen in my literature to come. If your data set contains hundred of observations an outlier or two may not cause, cause for alarm. However, outliers can spell trouble for models fitted to small data sets, since the sum of squares of the residuals is the basis for estimating parameters and calculating error statistics and confidence intervals, one or bad outliers in a small data set can badly skew the result. When outliers are found, two questions arise:

(a) Are they merely fluke of some kind? For instance data entry errors or the results of exceptional conditions that are not expected to recur or do they represent a real effect that you might want to include in your model.

(b) How much have the coefficients error statistics and predictions etc been affected?

An outlier may or may not have a dramatic effect on a model depending on the amount of "Leverage" that it has. Its leverage depends on the values of the independent variables at the point where it occurred. If the independent variables were all relatively close to their mean values, then the outliers may have a large influence in the estimate of the corresponding coefficients e.g. it may cause an otherwise insignificant variable to appear significant or vice versa. The best way to determine how much leverage on outlier (or group of outliers) has is to exclude it from fitting the model and compare the results with those originally obtained.

This paper is aimed at

(i) Checking a data set if it contains one or more observations that appear different from the rest of the data;

(ii) Checking the data value that does not conform to the remainder of the data;

(iii) If (ii) above is found, to check if the observation will cause the simple regression model to be substantially different from what they would be if the observation were removed from the data set;

5 SCOPE OF STUDY

The study is designed to check/detect outliers in a univariate, bivariate and multivariate data. Two univariate tests will be used: Rosner's and Grubbs. Influential observations will be checked in the bivariate data (simple linear regression) using data plots. Finally leverage value method will be used to detect outliers in a multiple regression model.

6 SIGNIFICANCE OF STUDY

This research will provide clues on the application of exploratory data analysis techniques that is involved in the detection of outliers in a univariate, bivariate and multivariate data and its evaluation on how they impact the results of an analysis, which if the contents are adequately understood by researchers will help to reach conclusions that are in line with their research objectives in their research works.

7 DEFINITION OF TERMS

- **Outlier**: An observation in which the studentized residual is large relative to other observations in the data set.
- **Influential observation**: An observation(s) that individually or jointly excessively influence the regression equation.
- **Fringelier**: Unusual events which occur more often than seldom.
- **Robust Method**: A statistical procedure to protect data from being distorted by outliers.
- **Mixture Regression**: This is a regression that occur when there is an omitted categorical predictor, thus regression occur in each category.
- **IQR Computation**: Inter-quartile range computation.

8 METHODOLOGY

This chapter was designed to explain the methods used by the researcher in analyzing data used in the study. Secondary data were used in this study.

The source of the data was from Dr. Kelly Uscategui, University of Connecticut on BOLLERS.DAT, Statistics, eighth edition. The data were collected through library research as shown in Appendix A.

9 REGRESSION ANALYSIS

Regression analysis is an estimating equation which expresses the functional relationship between two or more variables as well takes care of the error term which is classified into

1. Simple linear regression
2. Multiple linear regression

9.1 SIMPLE LINEAR REGRESSION

This is the type of linear regression that involves only two variables one independent and one dependent plus the random error term. The simple linear regression model assumes that there is a straight line (linear) relationship between the dependent variable y and the independent variable x. The model is expressed as

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad \ldots (1)$$

Y intercept β_0 and the slope β_1 are called the regression coefficients. The true value of the y-intercept (β_0) and slope (β_1) in the simple linear regression model are not known and can be estimated by the least square estimate methods and is expressed by

$$b_i = \frac{n\sum_{i=1}^{n} x_i y_i - \left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n\sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \quad \ldots (2)$$

where

n = number of observations

Then the least square point estimate of the y-intercept β_0 is expressed as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \ldots (3)$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \ldots (4)$$

To estimate how “good” our point estimates of β_0 and β_1 are, the visually fitted

line called the least square regression line or least square prediction equation is used. The general form of the equation is given by

$$\hat{y}_i = b_0 + b_{ix}$$

9.2 SCATTER DIAGRAM

In the analysis of the relationship between variables, it is often desirable to present a scatter diagram. A scatter diagram is used to investigate the pattern of the relationship between variables under investigation. It also gives a rough idea of how the variables x and y only are related and suggests the type of regression model to be fitted for the data. From there one can be able to determine if there is a positive or negative slope or if the points are curvilinear, exponential, non-linear or random.

For the purpose of this study, the scatter plots is used to detect an observation that is separated from the rest of the data called "outlier" this is done by plotting a data plot of the values of a dependent variable y against an independent variable x.

10 DATA ANALYSIS

We are analyzing data for 36 boilers collected for this research work. However, the statistical techniques discussed in this paper shall be used in this section.

IDENTIFICATION OF INFLUENTIAL OBSERVATION

We shall check for observations that cause the least squares point estimates to be substantially different from what they would be if the observation were removed from the data set (influential). The least square point estimates of the regression model $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ were as shown in Appendix B and written here as $\hat{Y} = -3870.3 + 0.00810 X_1 + 2.1592 X_2 + 3586.3 X_3 + 2020.3 X_4$ with R^2 = 0.930. The least square point estimates of the regression model when observations 4 and 19 were removed from the data as shown (in Appendix B, regression 2) as suspected outliers is $\hat{Y} = -1402 + 0.00717 X_1 + 0.8339 X_2 + 1960.5 X_3 + 2000.4 X_4$.

The coefficients of multiple determination R^2 of Y, X_1, X_2, X_3, X_4 was 0.930 which shows that 93.0% of the total variations of Y explained by the regression. In the second regression, only 86.3% of the total variations of Y explained by the regression. This implies that observations 4 and 19 are influential, because of the change in R^2.

An observation may be an outlier with respect of its Y value and or its X values but an outlier may or may not be influential. We can illustrate these ideas by considering a hypothetical plot of the values of the dependent variable Y against an independent variable X. We can also show this by considering a hypothetical plot of Y against X_1, Y against X_2, Y against X_3 and against X_4.

A plot of Y against X_1(see Appendix C,) shows that observations 4 and 19 are outlying with respect to both x value and y values observation 28(90000, 2635) is outlying with respect to x value. Also, the least square estimates of the regression model $\hat{Y} = \beta_0 + \beta_1 X_1$ is $\hat{Y} = 1760.3 + 0.00795 X_1$ (as shown in Appendix B, Regression 3) with R^2 = 0.685 and when observations 4 and 19 are removed, the least square estimates is $\hat{Y} = 2326.4 + 0.00535 X_1$ with R^2 = 0.467(Appendix A, Regression 4). We noticed again that the coefficient of multiple determination R^2 changes from 0.685 to 0.467. This also shows that observations 4 and 19 are influential.

A plot of Y against X_2 (see Appendix C, Graph 2) shows that observations 4 and 19 are outlying with respect to both x value and y values observation is outlying with respect to x value. Also, the least square estimates of the regression model $\hat{Y} = \beta_0 + \beta_1 X_1$ is

$\hat{Y} = 1859.6 + 3.1458X_2$ (as shown in Appendix B, Regression 3) with R^2 = 0.529 and when observations 4 and 19 are removed, the least square estimates is $\hat{Y} = 2869.1 + 1.3800X_2$ with R^2 = 0.131 (Appendix B, Regression 6). It obviously shows that observations 4 and 19 are seriously influential.

A plot of Y against X_3 (see Appendix C, Graph 3) shows that observations 4 and 19 are outlying with respect to y value. The least square estimates of the regression model $\hat{Y} = \beta_0 + \beta_1 X_3$ is $\hat{Y} = 7155.1 - 3682.8X_3$ with R^2 = 0.330 and when observations 4 and 19 are removed, the least square estimates is $\hat{Y} = 5270.8 - 1798.5X_3$ with R^2 = 0.170 (Appendix B, Regression 8). This shows that though observations 4 and 19 are outlying with respect to y value and not x values, the R^2 does not really differ.

A plot of Y against X_4 (see Appendix C, Graph 4) shows that observations 4 and 19 are outlying with respect to y value. The least square estimates of the regression model as shown in Appendix B, regression 9 is $\hat{Y} = 2783.8 - 2712.6X_3$ with R^2 = 0.256 and when observations 4 and 19 are removed, the least square estimates is $\hat{Y} = 2783.7 + 1900.2X_4$ with R^2 = 0.170 (Appendix B, Regression 10). This shows also that though observations 4 and 19 are outlying with respect to y value and not x values, the R^2 does not really differ. Thus, it is not influential.

11 SUMMARY, CONCLUSION AND RECOMMENDATION

11.1 SUMMARY

In identifying influential observation, the multiple regression model of Y on X_1, X_2, X_3, and X_4 showed that the least square estimate changed from when the two suspected outliers (observations 4 and 19) were removed. The R2 also reduced from 0.930 to 0.863. This shows that observations 4 and 19 are influential. A simple plot of Y on X_1 reveals that observations 4 and 19 are outlying with respect to both X value and Y values. The regression model of Y on X_1 showed that the least square estimate changed from when observation 4 and 19 were removed. The R^2 reduced from 0.685 to 0.467 which shows that observations 4 and 19 are influential.

A simple plot of Y on X_2 reveals that observations 4 and 19 are outlying with respect to both X and Y values. The R^2 reduced from 0.529 to 0.131 that shows that observations 4 and 19 are influential.

A simple plot of Y on X_3 reveals again that observations 4 and 19 are outlying with respect to Y values. The R^2 does not necessarily differ, thus observations 4 and 19 are not influential.

A simple plot of Y on X_4 reveals again that observations 4 and 19 are outlying with respect to Y values. The R^2 does not necessarily differ, thus observations 4 and 19 are not influential.

11.2 CONCLUSION

All of the above discussed statistical tests are used to determine if experimental observations are statistical outliers in the data set. If an observation is statistically determined to be an outlier this outlier before its exclusion is checked if it is influential. The observation should be treated as an extreme but valid measurement and it should be in further analysis.

Developing techniques to look for outliers and understanding how they impact data analysis are extremely important part of a thorough analysis, especially when statistical techniques are applied to the data. For example, in the procedure of outliers, any statistical test based on sample means and variances can be distorted. Estimated

regression coefficients that minimize the sum of squares for error (SSE) are very sensitive to outliers.

There are several other problematic effects of outliers including distortion of estimates, inflated sums of square you will be able to partition source of variation in the data into meaningful components, faulty conclusions, its quite possible to draw false conclusion if you have not looked for indication that there was anything usual in the data.

Effectively working with outliers in numerical data can be a rather difficult and frustrating experience. Neither ignoring nor deleting them at will is good solutions. If you do nothing, you will end up with a model that describes essentially none of the data, neither the bulk of the data nor the outliers. Even though your numbers may be perfectly legitimate, if they lie outside the verge of most of the data, they can cases potential computational and influence problems.

11.3 RECOMMENDATION

Having carried out this research work , the following recommendations are made;

1. We recommend that experimenters should keep good record for each experiment. All data should be recorded with any possible explanation or additional information.
2. We recommend that analyst should employ robust statistical methods. These methods are minimally affected by outliers.
3. If any observation is statistically determined to be an outlier, the analyst should determine an explanation for this outlier before exclusion from further analysis. If an explanation cannot be found, then the observation should be treated as an extreme but valid measurement and it should be in further analysis.
4. Finally, when analyst identifies outliers, he must decide what to do with it. Outliers that are obvious mistakes are corrected when possible, and the corrected values are inserted. If the correct value is not known and cannot be obtained, the datum might be excluded and statistical methods that were developed specifically for missing values situation could be used.

REFERENCES

Anscombe, F.J. (1960): Rejection of Outliers. Technometrics, 2, 123 – 147.

Atkinson, A.C. (1985): Plols, Transformation and Regression oxford, Oxford University press.

Atkinson, A.C. (1994): Fast Very Robust Methods for the Detection of Multiple Outliers Journal of the American.

Atkinson, A.C. and Riani, M. (2000): Robust Diagnostic Regression analysis New York, Stringer verlag.

Arthur J. Auwokeri, Practical Research Methodology, Design, Analysis and Reporting, 2nd edition.

Barnett. V. and Lewis T. (1984): Outliers in Statistical Data John Wiley and sons, New York.

Barnett. V. and Lewis T. (1994): Outliers in Statistical Data. 2nd ed., Chichester:Wiley.

Beckman R.J. and Cook R.D. (1983): "Outliers with Discussion Technometrics" Vol. 25, pp. 119 – 149.

Bening J.E. (2000). Asymptotic Theory of Testing Statistical Hypotheses Efficient Statistics Optimality Power Loss, and Deficiency, Dordrcht. VSP.

Berthouex P.M. and Brown L.C. (1994): Statistics for environmental engineers. CRC press, London.

Cook R.D. (1999): Regression Graphics Ideas for Studying Regressions through

Graphic New York: Wiley Statistics Improving Business Processes.

Cook R.D. and Weiberg S. (1982): "Residuals and Influence in Regression": Chapman and Hall. New York.

Cox D.R. and Hinkley D.V. (1974): Theoretical Statistics: London: Chapman and Hall.

Cook, R.D. and Weisberg S. (1982). "Residuals and Influence in Regression". Chapman and Hall: New York.

Dixon, W.J. (1950). Analysis of Extreme Values. Annuals of Mathematical Statistics, 21; 488 – 506.

Draper N.R. and John J.A. (1981): Influential Observations and Outliers in Regression.

Evans, V.P. (1999). Strategies for Detecting Outliers in Regression Analysis: An Introductory Primer.

Environment Protection Agency (1992). Statistical Training Course for Grand Water Monitoring Data Analysis EPA/530-R-93-003, Office of Solid Waste Washington DC.

Gibbons R.D. (1994). Statistical Methods for Groundwater Monitoring Van Nostrand Reinhold, New York.

Gilbert (1987): Gilbert, R.O. (1987). Statistical Methods for Environmental Pollution Monitoring Van Nostrand Rein hold, New York.

Grubbs F.E. (1950). Sample Criteria for Testing Outlying Observations. Annals of Mathematical Statistics.

Hadi A.S. and Simonoff J.S. (1993). Procedures for the identIfication of Multiple Outliers in Linear Models. Journals of American Statistical.

Hadi A.S. and Simonoff J.S. (1994). Improving the Estimate and Outlier Identification Properties of the Least Median of Squares and Minimum Volume Ellipsoid Estimators.

Hadi, (1992). "A Modification of a Method for the Detection of Outliers in Multivariate Samples", 1994 JRSSB 56:2, 393 – 396).

Hamilton, L.C. (1992). Regressions with Graphics: A second course in Applied Statistics. Monterey, CA: Brooks/Cole.

Hawkins, D.M. (1980). Identification of Outliers London: Chapman and Hall.

Hawkins D.M. (1983). Discussion of Paper of Beckman and Cook. Technometrics.

Huben P. (1981). Robust statistics, New York; Wiley

Hucks, S.W. (2000). Reading Statistics and Research (3rd ed.). New York: Longman.

Iglewicz B. and Hoaglin D.C. (1993). How to Detect and Handle Outliers. American Society for Quality Control M. Iwnkee, WI.

Iglewicz B. and Hoaglin D.C. (1993). How to Detect and Handle Outliers. Milwaukee, WI: ASQC Quality Press.

Jarrell, M.G. (1994). A Comparison of two procedures, the Mabalanobis Distiance and the Andrews – Pregibon Statistics, for identifying Multivariate Outliers. Researches in the Schools, 1:49 – 58.

Kitagawa G. and Akaike H. (1982). A quasi Bayesian Approach to Outlier Detection.

Kleinbaum D.G.; Kupper L.L.; Muller K.E. (1987). Applied Regression Analysis and Other Multivariate Methods. PWS-KENT Publishing Company, Boston.

Lane, K. (2002, February). What is robust Regression and how do you do it? Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

McClave J.T. and Sincich T. (2000). Statistics Eight Edition. Prentice Hall Upper Saddle River, New Jersey.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. The quantity Journal of

experimental psychology. 43(4), 907 – 912.

Nduka E.C. (1991). Principles of Applied Statistics I, 1st Edition, Crystal Publishers.

Orr, J.M.; Sackett, P.R. and Dubois, C.L.Z. (1991). Outlier Detection and Treatment in I/O Pschology: A Survey of Researcher Belief and an Empirical Illustration. Personnel Psychology, 44, 473 – 486.

Osborne, J.W. (2002). Notes on the use of Data Transformation. Practical Assessment, Research and Evaluation, 8, Available online at http://ericae.net|pare|getvn.asp?v=8&n=6.

Osborne, J.W.; Christiansen, W.R.I. and Gunter, J.S. (2001). Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of our field. A Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, W.A.

Sachs, L. (1982). Applied Statistics: A hand book of Techniques (2nd ed). New York: Springer Verlag.

Stevens, J.P. (1984). Outliers and Influential Points in Regression Analysis Psychological Bulletin, 95, 339 – 344.

Panofsky H.A. and Brie G.W. (1958). Some Applications of Statistics for Methodology, Pennsylan's State University.

Rosner's Multiple Outlier Test Technometrics 25, No.2 May, (1983), 165, 172.

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

Rousseeuw J.L. (1988). Evaluating Outlier Identification tests: Mahalanobis D Squared and Comrey, 23(2), 189 – 202.

Rees D.G. (1999). Essential Statistics. 4th edition, Chapman and Hall/CRC press.

Richard O. (1989). Statistical Methods for Environmental Pollution Monitoring.

Stuart A. and Ord K. J. (1987). Kendalls Advanced Theory of Statistics Vol. 15th edition. London Griffin.

Tabachnick, B.G. and Fidell, L.S. (2000). Using Multivariate Statistics, 4th edition.

Taylor J. (1987). Quality Assurance of Chem Measurements, Lewis Publishers, Chelsea M.I

Wainer, H. (1976). Robust Statistics: A Survey and some Prescriptions. Journal of Educational Statistics 1(4), 285 = 312.

Wilks D.S. (1995). Statistical Methods in the Atmospheric Sciences, Academic Press.

Zimmerman, D.W. (1994). A note on the Influence of Outliers on Parametric and Nonparametric Tests. Journals of General Psychology, 121(4), 391 – 401.

Zimmerman, D.W. (1995). Increasing the Power of Nonparametric Tests by Detecting and Downweighting Outliers. Journal of Experimental Education, 64(1), 71 – 78.

Zimmerman, D.W. (1998). Invalidation of Parametric Statistical Tests by Concurrent Violation of Two Assumptions. Journal of Experimental Education 67(1), 55 – 68.

APPENDIX A

Table 1: BOILERS DATA

S/N	Man-Hours	Boiler Capacity	Design Pressure	Boiler Type	Drum Type
1	3137	120000	375	1	1
2	3590	65000	750	1	1
3	4526	150000	500	1	1
4	10825	1073877	2170	0	1
5	4023	150000	325	1	1
6	7606	610000	1500	0	1
7	3748	88200	399	1	1
8	2972	88200	399	1	1
9	3163	88200	399	1	1

10	4065	90000	1140	1	1
11	2048	30000	325	1	1
12	6500	441000	410	1	1
13	5651	441000	410	1	1
14	6565	441000	410	1	1
15	6387	441000	410	1	1
16	6454	627000	1525	0	1
17	6928	610000	1500	0	1
18	4268	150000	500	1	1
19	14791	1089490	2970	0	1
20	2680	125000	750	1	1
21	2974	120000	375	1	0
22	1965	65000	750	1	0
23	2566	150000	500	1	0
24	1515	150000	250	1	0
25	2000	150000	500	1	0
26	2735	150000	325	1	0
27	3698	610000	1500	0	0
28	2635	90000	1140	1	0
29	1206	30000	325	1	0
30	3775	441000	410	1	0
31	3120	441000	410	1	0
32	4206	441000	410	1	0
33	4006	441000	410	1	0
34	3728	627000	1525	0	0
35	3211	610000	1500	0	0
36	1200	30000	325	1	0

Source: Dr. Kelly Uscategui, University of Connecticut.

Table 2: BOILERS DATA AS USED IN THE STUDY

	Man-Hours	**Boiler Capacity**	**Design Pressure**	**Boiler Type**	**Drum Type**
S/N	Y	X_1	X_2	X_3	X_4
1	3137	120000	375	1	1
2	3590	65000	750	1	1
3	4526	150000	500	1	1
4	10825	1073877	2170	0	1
5	4023	150000	325	1	1
6	7606	610000	1500	0	1
7	3748	88200	399	1	1
8	2972	88200	399	1	1
9	3163	88200	399	1	1
10	4065	90000	1140	1	1
11	2048	30000	325	1	1
12	6500	441000	410	1	1
13	5651	441000	410	1	1
14	6565	441000	410	1	1
15	6387	441000	410	1	1
16	6454	627000	1525	0	1
17	6928	610000	1500	0	1
18	4268	150000	500	1	1
19	14791	1089490	2970	0	1
20	2680	125000	750	1	1
21	2974	120000	375	1	0
22	1965	65000	750	1	0
23	2566	150000	500	1	0
24	1515	150000	250	1	0
25	2000	150000	500	1	0
26	2735	150000	325	1	0
27	3698	610000	1500	0	0
28	2635	90000	1140	1	0
29	1206	30000	325	1	0
30	3775	441000	410	1	0
31	3120	441000	410	1	0
32	4206	441000	410	1	0
33	4006	441000	410	1	0
34	3728	627000	1525	0	0
35	3211	610000	1500	0	0
36	1200	30000	325	1	0

NOTE: Y is the dependent variable which X_1, X_2, X_3 and X_4 are the independent variables.

For the purpose of this study, the following holds:

Y represents Man-Hours

X_1 represents Boiler Capacity

X_2 represent Design Pressure

X_3 represent Boiler Type

X_4 represent Drum Type

TABLE 3: BOILERS DATA OBSERVATION 4 AND 19 REMOVED

	Man-Hours	**Boiler Capacity**	**Design Pressure**	**Boiler Type**	**Drum Type**
S/N	Y	X_1	X_2	X_3	X_4
1	3137	120000	375	1	1
2	3590	65000	750	1	1
3	4526	150000	500	1	1
4	.	.	.	.	.
5	4023	150000	325	1	1
6	7606	610000	1500	0	1
7	3748	88200	399	1	1
8	2972	88200	399	1	1

9	3163	88200	399	1	1
10	4065	90000	1140	1	1
11	2048	30000	325	1	1
12	6500	441000	410	1	1
13	5651	441000	410	1	1
14	6565	441000	410	1	1
15	6387	441000	410	1	1
16	6454	627000	1525	0	1
17	6928	610000	1500	0	1
18	4268	150000	500	1	1
19	.	.	.	.	.
20	2680	125000	750	1	1
21	2974	120000	375	1	0
22	1965	65000	750	1	0
23	2566	150000	500	1	0
24	1515	150000	250	1	0
25	2000	150000	500	1	0
26	2735	150000	325	1	0
27	3698	610000	1500	0	0
28	2635	90000	1140	1	0
29	1206	30000	325	1	0
30	3775	441000	410	1	0
31	3120	441000	410	1	0
32	4206	441000	410	1	0
33	4006	441000	410	1	0
34	3728	627000	1525	0	0
35	3211	610000	1500	0	0
36	1200	30000	325	1	0

APPENDIX B

REGRESSION 1 (Complete data)

Results for: Worksheet 2

Regression Analysis: Y versus X1, X2, X3, X4

The regression equation is
Y = - 3870 + 0.00810 X1 + 2.16 X2 + 3586 X3 + 2020 X4

Predictor	Coef	SE Coef	T	P	VIF
Constant	-3870.3	861.2	-4.49	0.000	
X_1	0.0081030	0.0007875	10.29	0.000	2.982
X_2	2.1592	0.4382	4.93	0.000	4.551
X_3	3586.3	672.8	5.33	0.000	4.886
X_4	2020.3	259.9	7.77	0.000	1.042

S = 759.238 R-Sq = 93.0% R-Sq(adj) = 92.1%

PRESS = 32451614 R-Sq(pred) = 87.31%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	237794895	59448724	103.13	0.000
Residual Error	31	17869720	576443		
Lack of Fit	18	15796077	877560	5.50	0.002
Pure Error	13	2073643	159511		
Total	35	255664615			

15 rows with no replicates

Source	DF	Seq SS
X_1	1	175007141
X_2	1	4591993
X_3	1	23364628
X_4	1	34831133

Unusual Observations

Obs	X1	Y	Fit	SE Fit	Residual	St Resid
19	1089490	14791	13391	580	1400	2.86RX
20	125000	2680	4369	227	-1689	-2.33R
21	120000	2974	1498	204	1476	2.02R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 2.40696

REGRESSION 2 (Incomplete data)

Regression Analysis: Y versus X1, X2, X3, X4

The regression equation is
Y = - 1402 + 0.00717 X1 + 0.834 X2 + 1961 X3 + 2000 X4

34 cases used, 2 cases contain missing values

Predictor	**Coef**	**SE Coef**	**T**	**P**	**VIF**
Constant	**-1402**	**1164**	**-1.20**	**0.238**	
X_1	**0.0071706**	**0.0008296**	**8.64**	**0.000**	**2.377**
X_2	**0.8339**	**0.5930**	**1.41**	**0.170**	**5.132**
X_3	**1960.5**	**821.0**	**2.39**	**0.024**	**7.478**
X_4	**2000.4**	**229.9**	**8.70**	**0.000**	**1.005**

S = 667.350 R-Sq = 86.3% R-Sq(adj) = 84.4%

PRESS = 17963931 R-Sq(pred) = 80.93%

Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	**4**	**81263067**	**20315767**	**45.62**	**0.000**
Residual Error	**29**	**12915340**	**445357**		
Lack of Fit	**16**	**10841697**	**677606**	**4.25**	**0.006**
Pure Error	**13**	**2073643**	**159511**		
Total	**33**	**94178407**			

13 rows with no replicates

Source	DF	Seq SS
X_1	1	44016857
X_2	1	251199
X_3	1	3267135
X_4	1	33727876

Unusual Observations

Obs	X1	Y	Fit	SE Fit	Residual	St Resid
6	610000	7606	6223	296	1383	2.31R
20	125000	2680	4081	221	-1401	-2.22R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.18328

REGRESSION 3 (Complete data)

Regression Analysis: Y versus X1

The regression equation is
Y = 1760 + 0.00795 X1

Predictor	Coef	SE Coef	T	P	VIF
Constant	1760.3	390.8	4.50	0.000	
X_1	0.0079456	0.0009251	8.59	0.000	1.000

S = 1540.22 R-Sq = 68.5% R-Sq(adj) = 67.5%

PRESS = 101555418 R-Sq(pred) = 60.28%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	175007141	175007141	73.77	0.000
Residual Error	34	80657474	2372279		
Lack of Fit	10	36814938	3681494	2.02	0.078
Pure Error	24	43842536	1826772		
Total	35	255664615			

3 rows with no replicates

Unusual Observations

Obs	X_1	Y	Fit	SE Fit	Residual	St Resid
4	1073877	10825	10293	744	532	0.39 X
19	1089490	14791	10417	758	4374	3.26RX
34	627000	3728	6742	384	-3014	-2.02R
35	610000	3211	6607	372	-3396	-2.27R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 0.912587

REGRESSION 4 (Incomplete data)

Regression Analysis: Y versus X1

The regression equation is
Y = 2326 + 0.00535 X1

34 cases used, 2 cases contain missing values

Predictor	Coef	SE Coef	T	P	VIF
Constant	2326.4	349.8	6.65	0.000	
X_1	0.005349	0.001009	5.30	0.000	1.000

S = 1252.02 R-Sq = 46.7% R-Sq(adj) = 45.1%

PRESS = 58199599 R-Sq(pred) = 38.20%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	44016857	44016857	28.08	0.000
Residual Error	32	50161550	1567548		
Lack of Fit	8	6319014	789877	0.43	0.890
Pure Error	24	43842536	1826772		
Total	33	94178407			

1 rows with no replicates

Unusual Observations

Obs	X_1	Y	Fit	SE Fit	Residual	St Resid
35	610000	3211	5589	402	-2378	-2.01R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 0.627407

REGRESSION 5 (Complete data)

Regression Analysis: Y versus X2

The regression equation is
Y = 1860 + 3.15 X2

Predictor	Coef	SE Coef	T	P	VIF
Constant	1859.6	503.4	3.69	0.001	
X_2	3.1458	0.5093	6.18	0.000	1.000

S = 1882.45 R-Sq = 52.9% R-Sq(adj) = 51.5%

PRESS = 153454684 R-Sq(pred) = 39.98%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	135181595	135181595	38.15	0.000
Residual Error	34	120483020	3543618		
Lack of Fit	10	75225613	7522561	3.99	0.003
Pure Error	24	45257407	1885725		
Total	35	255664615			

3 rows with no replicates

Unusual Observations

Obs	X_2	Y	Fit	SE Fit	Residual	St Resid	
4	2170	10825	8686	778	2139	1.25	X
19	2970	14791	11203	1162	3588	2.42	RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 0.978152

REGRESSION 6 (Incomplete data)

Regression Analysis: Y versus X2

The regression equation is
Y = 2869 + 1.38 X2

34 cases used, 2 cases contain missing values

Predictor	Coef	SE Coef	T	P	VIF
Constant	2869.1	500.2	5.74	0.000	
X_2	1.3800	0.6271	2.20	0.035	1.000

S = 1598.81 R-Sq = 13.1% R-Sq(adj) = 10.4%

PRESS = 94041398 R-Sq(pred) = 0.15%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	12380689	12380689	4.84	0.035
Residual Error	32	81797718	2556179		
Lack of Fit	8	36540311	4567539	2.42	0.045
Pure Error	24	45257407	1885725		
Total	33	94178407			

1 rows with no replicates

Durbin-Watson statistic = 0.711541

REGRESSION 7 (Incomplete data)

Regression Analysis: Y versus X_3

The regression equation is
Y = 7155 - 3683 X_3

Predictor	Coef	SE Coef	T	P	VIF
Constant	7155.1	793.5	9.02	0.000	
X_3	-3682.8	899.8	-4.09	0.000	1.000

S = 2244.43 R-Sq = 33.0% R-Sq(adj) = 31.0%

PRESS = 209979555 R-Sq(pred) = 17.87%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	84390625	84390625	16.75	0.000
Residual Error	34	171273989	5037470		
Total	35	255664615			

The number of distinct predictor combinations equals the number of
parameters.
No degrees of freedom for lack of fit.
Cannot do pure error test.

Unusual Observations

Obs	X_3	Y	Fit	SE Fit	Residual	St Resid	
19	0.00	14791	7155	794	7636	3.64	R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.30457

REGRESSION 8 (Complete data)

Regression Analysis: Y versus X_3

The regression equation is
Y = 5271 - 1798 X_3

34 cases used, 2 cases contain missing values

Predictor	Coef	SE Coef	T	P	VIF
Constant	5270.8	638.2	8.26	0.000	
X_3	-1798.5	703.2	-2.56	0.015	1.000

S = 1563.21 R-Sq = 17.0% R-Sq(adj) = 14.4%

PRESS = 90911413 R-Sq(pred) = 3.47%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15982317	15982317	6.54	0.015
Residual Error	32	78196089	2443628		

Total 33 94178407

The number of distinct predictor combinations equals the number of parameters.
No degrees of freedom for lack of fit.
Cannot do pure error test.

Unusual Observations

Obs	X_3	Y	Fit	SE Fit	Residual	St Resid
14	1.00	6565	3472	295	3093	2.01R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 0.788722

REGRESSION 9 (Incomplete data)

Regression Analysis: Y versus X_4

The regression equation is
$Y = 2784 + 2713\ X_4$

Predictor	Coef	SE Coef	T	P	VIF
Constant	2783.8	591.4	4.71	0.000	
X_4	2712.6	793.4	3.42	0.002	1.000

S = 2365.55 R-Sq = 25.6% R-Sq(adj) = 23.4%

PRESS = 211244596 R-Sq(pred) = 17.37%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	65406211	65406211	11.69	0.002
Residual Error	34	190258404	5595835		
Total	35	255664615			

The number of distinct predictor combinations equals the number of parameters.
No degrees of freedom for lack of fit.
Cannot do pure error test.

Unusual Observations

Obs	X_4	Y	Fit	SE Fit	Residual	St Resid
4	1.00	10825	5496	529	5329	2.31R
19	1.00	14791	5496	529	9295	4.03R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.28861

REGRESSION 10 (Complete data)

Regression Analysis: Y versus X_4

The regression equation is
$Y = 2784 + 1900\ X_4$

34 cases used, 2 cases contain missing values

Predictor	Coef	SE Coef	T	P	VIF
Constant	2783.7	352.4	7.90	0.000	
X_4	1900.2	484.4	3.92	0.000	1.000

S = 1409.71 R-Sq = 32.5% R-Sq(adj) = 30.4%

PRESS = 71537033 R-Sq(pred) = 24.04%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	30585083	30585083	15.39	0.000
Residual Error	32	63593324	1987291		
Total	33	94178407			

The number of distinct predictor combinations equals the number of parameters.
No degrees of freedom for lack of fit.
Cannot do pure error test.

Unusual Observations

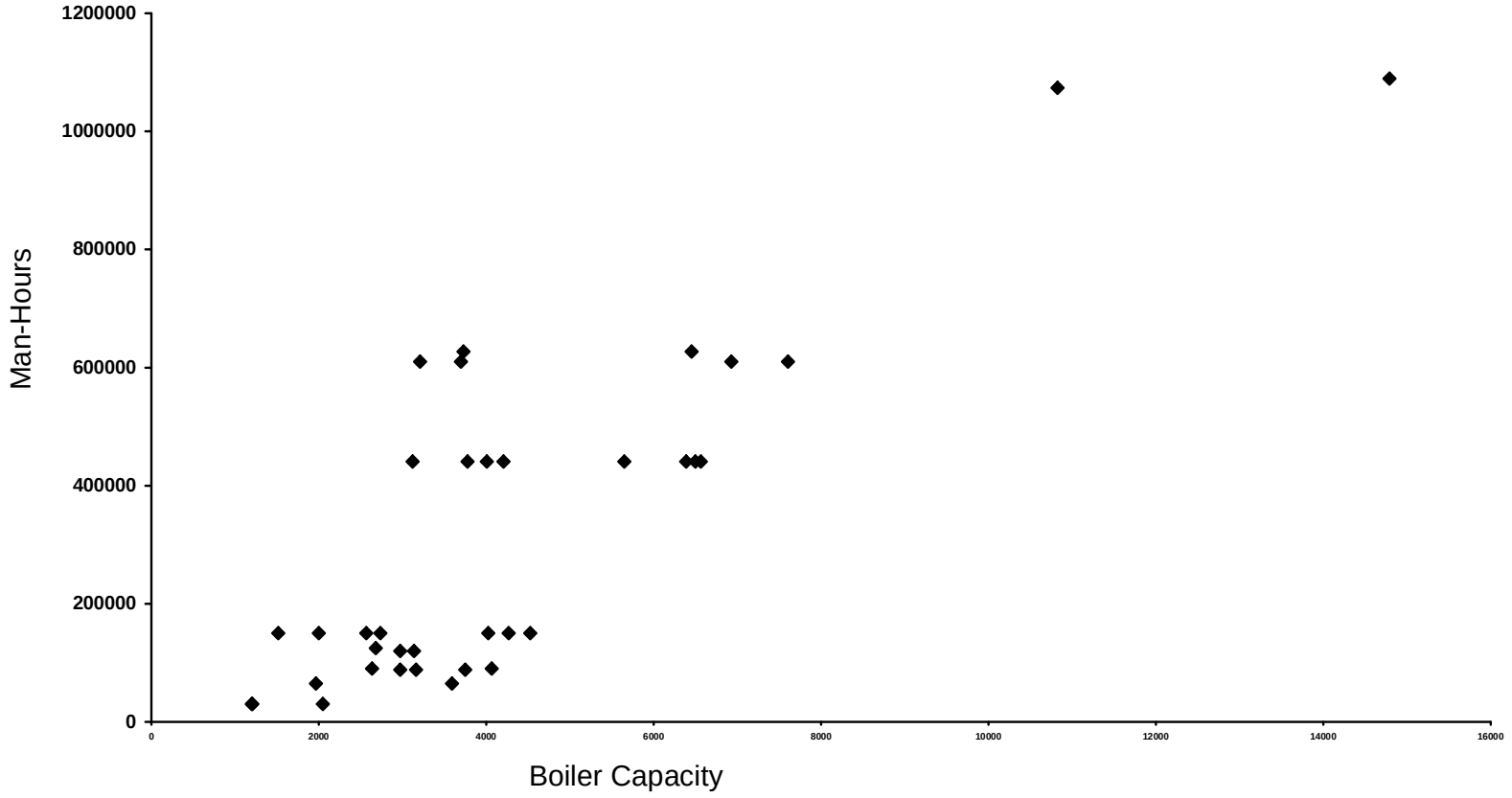
Obs	X_4	Y	Fit	SE Fit	Residual	St Resid
6	1.00	7606	4684	332	2922	2.13R

R denotes an observation with a large standardized residual.

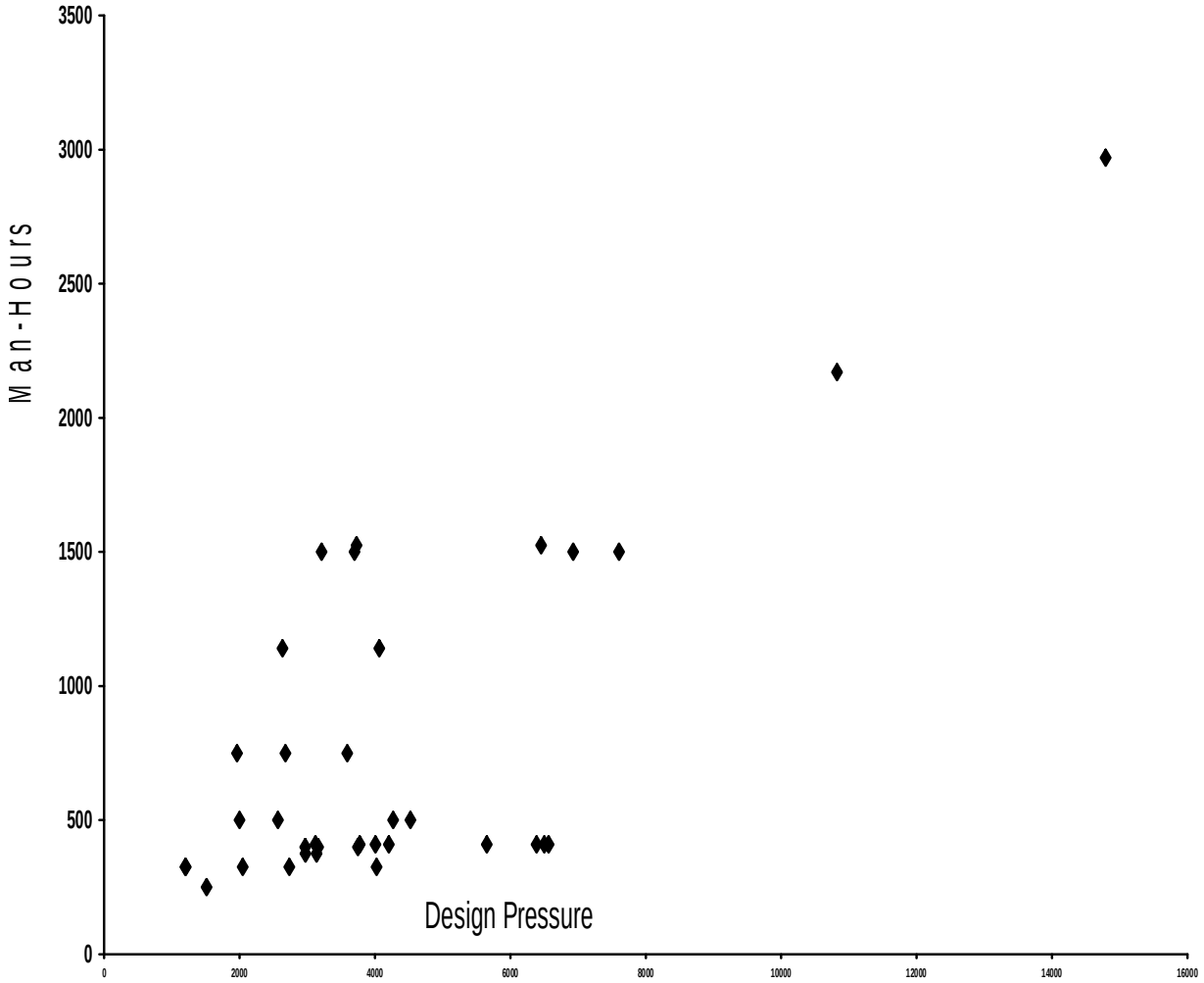
Durbin-Watson statistic = 1.42502

APPENDIX C

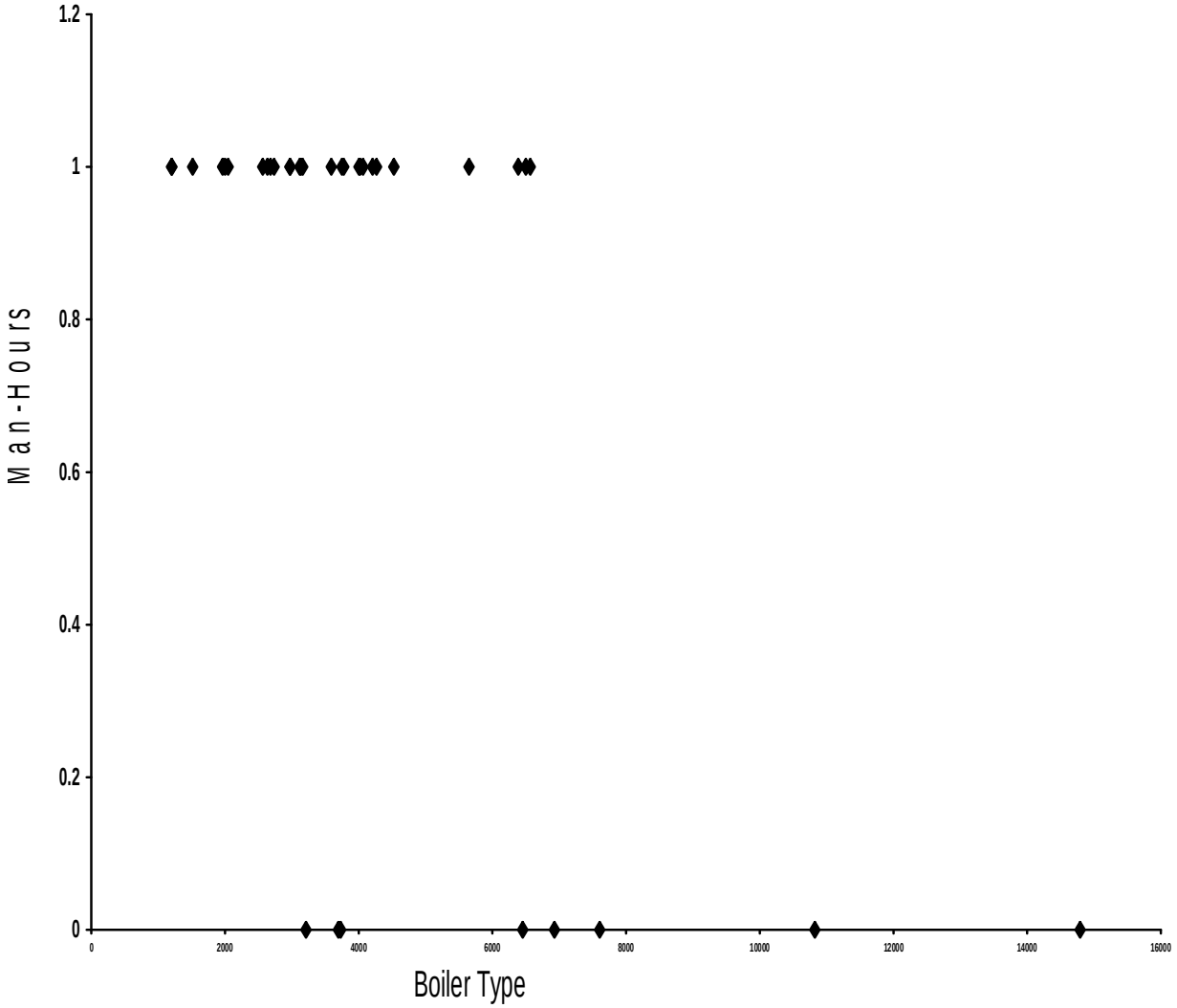
Graph 1



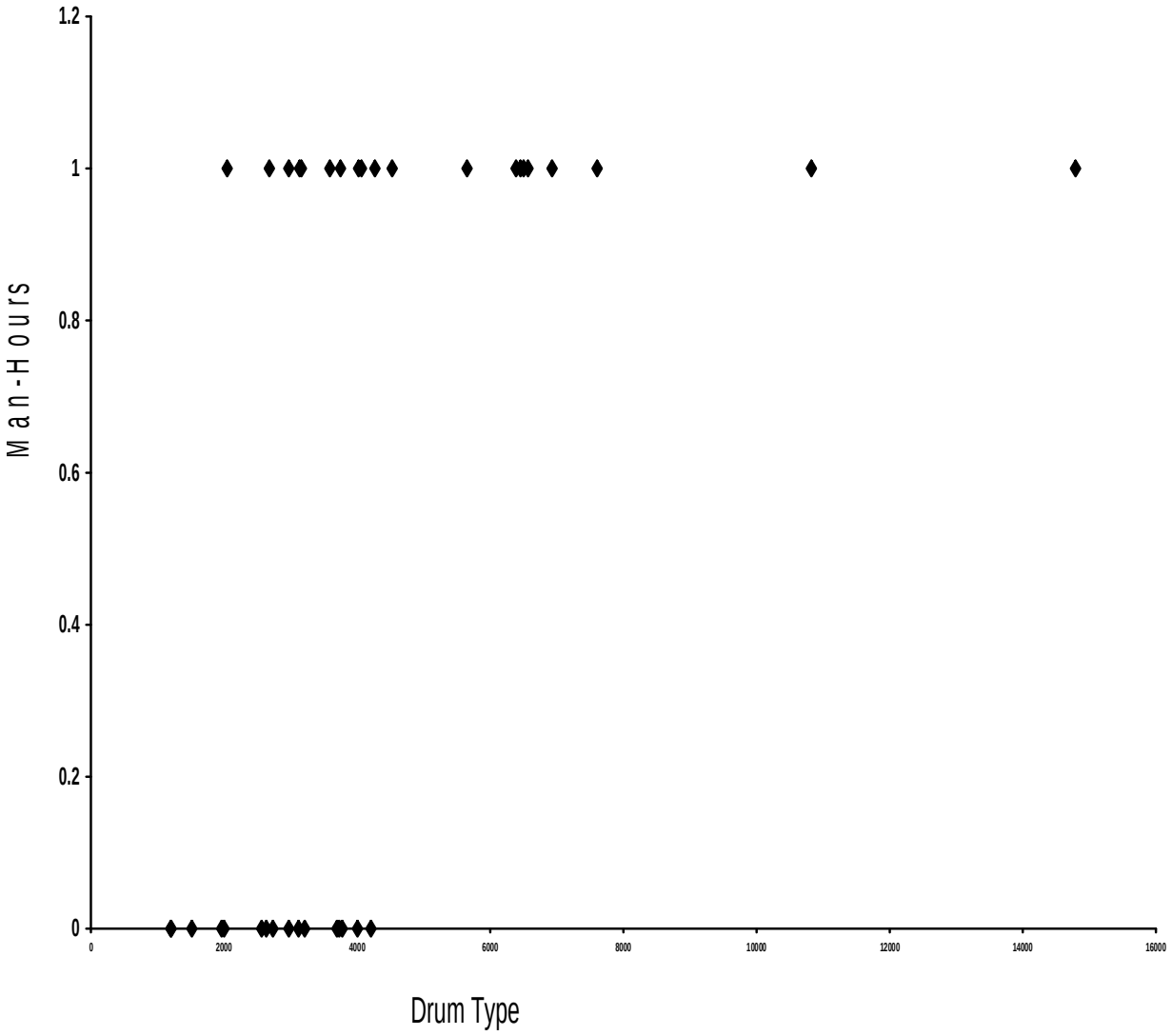
Graph 2


3500
3000
2500
2000
1500
1000
500
0
Man-Hours
Design Pressure
0
2000
4000
6000
8000
10000
12000
14000
16000

Graph 3


1.2
1
0.8
0.6
0.4
0.2
0
Man-Hours
0
2000
4000
6000
8000
10000
12000
14000
16000
Boiler Type

Graph 4


1.2
1
0.8
0.6
0.4
0.2
0
Man-Hours
0
2000
4000
6000
8000
10000
12000
14000
16000
Drum Type